The Problem of Extrapolation in Economics

Sofia Alexandra Blanco Sequeiros

UNIVERSITY OF HELSINKI

Faculty of Social Sciences

Practical Philosophy

Master's thesis

May, 2019

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Social Sciences	Department of Philosophy

Tekijä — Författare — Author
Sofia Alexandra Blanco Sequeiros

Työn nimi — Arbetets titel — Title
The Problem of Extrapolation in Economics

Oppiaine — Läroämne — Subject
Practical Philosophy

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	May 6, 2019	86 pages

Tiivistelmä — Referat — Abstract

This thesis explores the problem of extrapolating causal claims in the social sciences, particularly economics. The problem of extrapolation is the problem of inferring something about a phenomenon of interest in one context, based on what is known about it in another. For example, we may want to infer that a medicine works in population Y, based on the fact that we know it works in population X. Extrapolation is the inferential process of generalizing or transporting claims about a phenomenon of interest to new populations or settings. The answers to the problem of extrapolation in philosophy of science aim to explain how successful extrapolation is possible, as there will always be relevant differences between the two systems.

I study extrapolation from the viewpoint of philosophy of science, which aims to both analyze and complement science and scientific knowledge. I also use a case study with two examples to further illustrate the relationship between the theoretical approaches to extrapolation in philosophy of economics and actual studies in experimental economics. I focus on comparative process tracing, a general account of extrapolation developed by philosopher of science Daniel Steel, and its success in extrapolating causal claims from field experiments in economics.

The first chapter introduces central concepts and key questions. The second chapter discusses external validity, a concept typically used in economics to describe the potential of causal claims to be extrapolated. The third chapter introduces comparative process tracing, which explains how and why extrapolation can be based on knowledge about causal mechanisms. Next, I discuss field experiments in economics and methodological issues of extrapolation particular to them. The fourth chapter consists of a case study, which shows the limitations of approaching extrapolation in economics with comparative process tracing. The last chapter concludes.

The central conclusion of this thesis is that even though comparative process tracing is meant as an account of extrapolation that can explain and apply to extrapolation across disciplines, applying it to economics faces methodological challenges. Nevertheless, the issues it faces with regard to field experiments in economics do not refute it as an account of mechanistic extrapolation. I propose that comparative process tracing is a theoretically comprehensive epistemological account of extrapolation in the social sciences, but it must be complemented with a systematic methodological account of problems of extrapolation in practice. This methodological account complements and enhances epistemological analysis of extrapolation.

— Nyckelord — Keywords
extrapolation, field experiments, philosophy of economics, philosophy of science

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Valtiotieteellinen tiedekunta	Käytännöllisen filosofian laitos

Tekijä — Författare — Author
Sofia Alexandra Blanco Sequeiros

Työn nimi — Arbetets titel — Title

The Problem of Extrapolation in Economics

Oppiaine — Läroämne — Subject
Käytännöllinen filosofia

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Maisterintutkielma	6.5.2019	86 sivua

Tiivistelmä — Referat — Abstract

Tutkielma tarkastelee ekstrapolointia eli yleistämistä yhteiskuntatieteissä. Se keskittyy kausaalisten johtopäätösten yleistämiseen taloustieteessä. Ekstrapolointi on ongelma tilanteissa, joissa halutaan päätellä jotain tutkittavasta ilmiöstä yhdessä populaatiossa tai ympäristössä sen perusteella, mitä siitä tiedetään toisaalla. Yksinkertaistaen: voidaan päätellä, että lääke toimii populaatiolla Y sen perusteella, että lääkkeen tiedetään toimivan populaatiolla X . Ekstrapolointi on päättelyprosessi, jossa tutkittavaa ilmiötä koskevat kausaaliset väitteet yleistetään tai siirretään uusiin populaatioihin tai ympäristöihin. Tieteenfilosofiset vastaukset ekstrapoloinnin ongelmaan pyrkivät selittämään, miten ekstrapoloinnin onnistuminen on mahdollista, vaikka populaatioiden ja ympäristöjen välillä on aina eroja.

Tutkin ekstrapolointia tieteenfilosofian menetelmin. Tieteenfilosofisen tutkimuksen tavoitteena on tutkia ja täydentää tieteellistä tietoa. Keskityn tutkielmassani tieteenfilosofi Daniel Steelin kehittämään, vertailevaksi prosessinseurannaksi kutsuttuun lähestymistapaan, jonka mukaan ekstrapolointi voi yhteiskuntatieteissä perustua tietoon kausaalisista mekanismeista. Erityisesti keskityn siihen, kuinka hyvin vertaileva prosessinseuranta vastaa taloustieteen kenttäkokeissa nouseviin ekstrapoloinnin ongelmiin.

Ensimmäinen luku esittelee avainkäsitteet ja -kysymykset. Toinen luku tarkastelee ulkoisen validiteetin käsitettä, jota käytetään taloustieteessä kuvaamaan kausaalisten johtopäätösten yleistettävyyttä. Kolmas luku esittelee vertailevan prosessinseurannan keskeiset periaatteet. Seuraavaksi tutkin taloustieteen kenttäkokeita ja nistä yleistämiseen liittyviä menetelmällisiä kysymyksiä ja ongelmia. Neljäs luku tarkastelee kahta esimerkkitapausta, jotka osoittavat vertailevan prosessinseurannan menetelmälliset rajoitteet taloustieteessä. Viimeinen luku päättää.

Keskeinen johtopäätös on, että vaikka vertailevan prosessinseurannan on tarkoitus vastata ekstrapoloinnin ongelmiin monilla tieteenaloilla, sen soveltaminen taloustieteeseen on vaikeaa. Menetelmälliset haasteet, joita sen soveltamisesta nousee, eivät kuitenkaan kumoa vertailevaa prosessinseurantaa mekanismeihin perustuvan ekstrapoloinnin teoriana. Esitän, että vertaileva prosessinseuranta on teoreettisesti kattava lähestymistapa ekstrapolointiin myös yhteiskuntatieteissä, mutta sitä tulee täydentää systemaattisella katsannolla menetelmällisiin ongelmiin, joita ekstrapolointi kohtaa käytännössä. Ymmärrys menetelmällisistä ongelmista täydentää ekstrapoloinnille keskeisten epistemologisten ongelmien analyysia.

— Nyckelord — Keywords
ekstrapolointi, kenttäkokeet, taloustieteen filosofia, tieteenfilosofia

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction

1.1 The Problem of Extrapolation

Experiments are tools for scientific inquiry. They are a systematic way of studying causality and identifying causal relationships. An experiment implements an intervention on a putative causal relationship: it manipulates an assumed causal factor X in order to see how the intervention affects the outcome, Y. Ideally, the effects of any factors that might disturb the causal relationship are controlled. (Guala, 1999, p. 555) Depending on the researcher's aim, experiments can be used for a variety of purposes, including theory testing and data collection. Experimental results can be used to draw inferences about the studied phenomenon in order to clarify its nature, explain why it happens, or predict whether or how it will happen in the future.

Experiments can be used as surrogate systems of reasoning when a phenomenon of interest cannot be directly studied for one reason or another (Baetu, 2015, p. 947). This is also the case with economic behavior, which can often be observed but not studied directly. We can observe that banks give small business loans to applicants who are in different situations financially, but it is impossible to observe all business loans given by all banks in real time. We cannot determine the precise effect of a loan on a person's financial situation just by observing the loans' effects in a scattered bunch of individual recipients.

Without controlled study, we cannot determine whether any positive effects came about because of the loan, or because at the same time, the applicants won in the lottery, inherited money, or completed a degree that put them in a different position on the job market. Thus, experimental economists have devised ways of using experiments to investigate the effects that microcredit programs, where small loans are given to applicants, have on poverty. In the experiments, microcredit is

given to some participants but not to others. The effects of the given microcredit are then systematically tracked within the two groups and the results compared. In other words, economists use a separate, smaller, controlled system to figure what could be going on in more complex systems.

The idea of experiments as surrogate systems of reasoning is to create a representative system from which inferences about a phenomenon can be generalized (Baetu, 2015). Inferences about the phenomenon in an experimental system have to be *extrapolated*, if they are taken as correct in or applied to somewhere else than the experimental system. The problem, especially in the social sciences, is that the causal relationships that hold in the small, simplified experimental system may not hold in other, more complex circumstances. Observing the effects of a microcredit program in the smaller system is easier, but in the more complex system, people, businesses and banks might act, and their behavior be interpreted, differently. Because of this, inferences from an experimental system to a target circumstance or population are always inductive.

This is the problem of extrapolation, which is the focus of this thesis. The problem of extrapolation is a concern for both observational and experimental studies in a variety of fields, including the social sciences, medicine, biology, psychology and climate science (see e.g. Guala, 2005; Steel, 2008; Parker, 2010; Cartwright & Hardie, 2012; Baetu, 2015; Westreich et al., 2018). At its most general, extrapolation is the process or activity of inductively inferring conclusions about something, for example a causal effect, in one situation or system, based on inferences about the same (or a similar) causal effect in another system (Steel, 2008, pp. 3, 78-79, 87-89). Westreich et al. distinguish between two kinds of extrapolation: extrapolation as generalization, which means generalizing causal claims about an effect observed in a study sample to the population that the study sample is drawn from. Extrapolation as transportation means transporting the claims to entirely new populations or con-

texts. The concepts of "generalizability" and "transportability" correspond to these. (Westreich et al., 2018, p. 438) This is a distinction I will follow in this thesis.

Experimental science, climate models, clinical trials, and policy recommendations use extrapolation in order to explain and predict causal effects within different systems. Issues of extrapolation are of philosophical and scientific, conceptual and methodological interest. Issues of extrapolation are particularly pertinent if experiments are used to guide policy. If economic experiments are to be used as a source of evidence in successful policy-making, inferences from them have to be applicable outside the experimental system (Reiss, 2013, p. 187). Extrapolating an inference about the effectiveness of a policy from one group of people to another has consequences not only for the people included in the study, but everyone included in policy recommendations drafted on the basis of research. Overall, extrapolation is a key issue in applying inferences about the results of scientific studies to new contexts (cf. Guala, 2005; Steel, 2008; Cartwright & Hardie, 2012; Baetu, 2015). It also intersects with a variety of questions in philosophy of science, such as causation, causal inference, causal explanation, and causal heterogeneity; prediction; the role of models, simulations and experiments in science, and their epistemology and methodology; and evidence and expertise (cf. Steel, 2008).

Extrapolation is a multidimensional question related to a variety aspects of scientific inquiry, both theoretical and practical. It is a concern for many different fields of science, and there are many kinds of extrapolation: from an experimental system to a non-experimental system, from one population to another, from a population to an individual, from an individual to an individual, and from an individual to a population. You can extrapolate in time, between geographical locations, or between species. Many kinds of evidence, theoretical and empirical, can be used to justify extrapolation, and sometimes extrapolation can succeed based only on a lucky guess. Different kinds of inferences and claims can be extrapolated, including claims about

causal effects, about correlations, and about characteristics or properties.

Extrapolation may be necessary, such as in cases where an experiment is conducted to inform and design economic or health policies. In those cases, it is crucial that the experimental system represent the non-experimental system, and that the extrapolation from the former to the latter be justified. Extrapolation can also be more speculative, such as in cases where researchers conduct an experiment and discuss whether those experiments tell us about the same phenomenon in other circumstances. In these cases, extrapolation can take on many forms, from vague speculation to rigorous use of experimental evidence for explaining or predicting phenomena in new domains.

This multidimensionality alone poses a challenge for any general investigation of extrapolation, and that is why this thesis focuses on the extrapolation of causal claims from experimental systems to non-experimental targets. I focus on the extrapolation of claims about causal effects and relationships, because that is what the philosophical literature on extrapolation has focused on thus far, and that is what the most developed accounts of extrapolation concern. Policy-relevant studies in economics, especially studies that aim at evaluating policies, are typically interested in finding causal effects. Consequently, it is of interest whether claims about them can be applied to new contexts.

This thesis focuses particularly on comparative process tracing, a proposed solution to the problem of extrapolation in social science and biology. Comparative process tracing is an account of extrapolation that grounds warranted extrapolation in the similarity of the causal mechanisms in the system where the causal claims are extrapolated from and the system they are extrapolated to (Steel, 2008). The account was developed by Steel (2008), and Steel presents it as a solution to two epistemological puzzles regarding the knowledge one needs for extrapolation to be

both successful and justified[1]. However, comparative process tracing is not meant only relevant in theory and for philosophy, but also in practice for scientific methodology (Steel, 2008, p. 6). Steel uses an example from biology, about the carcinogenic effects of aflatoxin B1, to show that comparative process tracing solves the epistemological challenges of extrapolation, and works as an account of extrapolation in practice (cf. Steel, 2008, pp. 79-80, 88-96).

In addition to biology, comparative process tracing is meant as an account of extrapolation that can be "usefully employed in social science", and in principle, there is no reason it could not be (Ibid, 149-150). Nevertheless, Steel is skeptical about comparative process tracing succeeding in the social sciences (Ibid). This thesis takes a look at whether Steel's skepticism regarding extrapolation as comparative process tracing in the social sciences is warranted. Philosophers of economics have discussed both the epistemic and the methodological problems that extrapolating causal claims from experiments in economics faces, but the discussion on other than laboratory experiments and experiments studying behavior in naturally occurring contexts is very small (cf. Viceisza 2016, 386). This thesis complements that discussion by exploring the consequences of using comparative process tracing to approach extrapolation in other kinds of field experiments in economics. Drawing on the extant literature, I give an overview of the methodological challenges that applying comparative process tracing to economics faces.

Baetu (2015) writes on extrapolation in the life sciences. He analyzes how scientists think about extrapolation, and argues that they treat it as an issue of "taking an epistemic risk" (Baetu, 2015, p. 944). According to him, thinking about a universally applicable solution to extrapolation is misguided. Instead, one should ask how scientific research can be planned so that the possibility of error in taking that epistemic risk is controlled. (Ibid.) The question underlying the analysis of this

[1]These will be discussed in more detail in chapter three.

thesis is whether Baetu's arguments also apply to extrapolation in the social sciences, particularly economics. Overall, they are supported by my study of extrapolation in economics.

1.2 Field Experiments

Field experiments are tied to a variety of theoretical and methodological questions related extrapolation. As the case study will show, field experiments exemplify the need to supplement comparative process tracing with a more specific account of the methodological questions related to not only external validity, but also extrapolation. Economists use laboratory, field and natural experiments to study behavior, decision-making and economic phenomena (Reiss, 2013, p. 174-175, 192-193). In laboratory and field experiments, the experimenter studies causality by manipulating causal factors one at a time, controlling possible confounders, and observing the outcomes of this iterative process (Guala, 1999, p. 555-556). In natural experiments, nature or society provides the experimental intervention and the researcher only needs to gather, organize and interpret data (Harrison & List, 2004, p. 1011, 1041-1042). In field experiments, the context is less artificial or strictly controlled as a laboratory procedure, but the aim is nonetheless to keep control over some or all of the putative causal factors (Harrison & List, 2004).

Field experiments can and are used for a wide variety of practical purposes, from testing theories to informing policy. Laboratory experiments in economics can be divided into three categories: those that "Speak to Theorists", "Search for Facts" and "Whisper into the Ears of Princes" (Roth, 1995, p. 22). Experiments in the first category are ones that empirically test formal theories and investigate unobserved regularities and anomalies. Experiments that search for facts are conducted in order to collect information about the phenomenon of interest. Whispering into the ears

of princes is the task of experiments that aim at guiding policy and showing policy-makers effective ways of reaching policy goals. (Ibid.) Field experiments also fit into these categories (List, 2007, p. 2). This thesis is especially interested in experiments that aim to provide policy-relevant results, which often fall into more than one of these categories.

Field experiments are conducted for a wide variety of reasons, methodological and otherwise. Large-scale field experiments enable economists to study and gather a large amount of data about different kinds of phenomena, like taxes, health and employment. Field experiments may provide a solution to problems that the presumed artificiality of laboratory experiments can cause with respect to extrapolation (Jimenez-Buedo & Miller, 2010; Jimenez-Buedo & Guala, 2016). Third, field experiments can be conducted to study the limitations of laboratory experiments (cf. Levitt & List, 2007a,b). By conducting the same experiment with the same participants in the laboratory and in the field, List (2006) is able to study how well laboratory results extrapolate to the field. Others doubt the inherent generalizability and transportability of inferences from field experiments for various reasons, for example because they might turn out to be good for studying very situation-specific causal effects, which do not extrapolate well (Gneezy & Imas, 2017, 440).

1.3 Method and Structure of the Thesis

This thesis studies extrapolation mainly within the context of experimental science, through theoretical frameworks provided by philosophy of science. According to Steel (2008), his book "*explores how and under what circumstances reliable extrapolation is possible in biology and social science, and explores some of the implications of this topic for issues in philosophy of biology and social science.*" (Steel, 2008, p. 4). After Steel's book, much of the development of accounts of extrapolation

in philosophy of science has happened through discussion of the methodological aspects of extrapolation, in addition to discussing the epistemological questions that extrapolation raises. This thesis continues the aforementioned extrapolation and the ongoing discussions.

Chang (2004) writes about history and philosophy of science as *complementary science*, that is, fields of inquiry whose aims are continuous with science itself. History and philosophy of science complement science by generating scientific knowledge regarding the questions that the sciences (what Chang calls "special sciences") have no resources to focus on, or need to neglect out of necessity to specialize. History and philosophy of science do not include science, but they can analyze science (Ibid, 236-240). Philosophy provides conceptual tools for "organized skepticism and criticism", and history is a central supply of the forgotten questions and answers of science (Ibid, 240) Ultimately, the aim is to study not only science, but that which science studies: nature (Ibid, 237).

This is also the methodological starting point of this thesis. In developing comparative process tracing, Steel's aim is not only to understand scientific practice, but also the nature of causality, extrapolation and the phenomena different sciences investigate. He analyzes extrapolation conceptually and methodologically, both as a process of inductive reasoning about causality as well as a practice in science. By doing so, he explains how, when and why extrapolation in biology and the social sciences can be justified. In order to investigate comparative process tracing as an account of extrapolation in the social sciences, I follow the methods articulated by Chang and put into use by Steel. I also conduct a case study to take a closer look at how comparative process tracing can be applied to actual cases in economics.

The use of case studies in philosophy of science is not uncontested (cf. Currie, 2015). Science is heterogeneous, and philosophers tend to use particular case studies to draw generalizations about science in an epistemically unjustified way (Ibid, p.

3). However, I use case studies not to draw inductive generalizations about science, but to show a philosophical point. Comparative process tracing is a theoretically comprehensive account of extrapolation and remains important because of its contribution to understanding the importance of causal mechanisms and information about them in extrapolation. With the case study, I conclude that in order for comparative process tracing to fully account for extrapolation in economics, the practical challenges related to experimental methodology have to be taken into account also by epistemological approaches to extrapolation. To understand extrapolation as a philosophical and a scientific issue, comparative process tracing has to be complemented with a methodological account detailing how to address the challenges that extrapolating with mechanisms faces in practice.

In the second chapter, I introduce the notion of external validity, with which problems of extrapolation are typically conceptualized in economics. Next, I present comparative process tracing. The general idea in process tracing is to trace the mechanisms in the model circumstance and combine that information with background knowledge about mechanisms in the target population to conclude whether or not it is likely that the causal claim drawn on the basis of the model will hold in the target (Steel, 2008, pp. 7-8). After presenting comparative process tracing, I discuss field experiments, and the issues that are relevant to extrapolation and external validity regarding their methodology. The fourth chapter consists of the case study, and the fifth chapter concludes.

2 Internal and External Validity

Experiments are correctly performed when experimenters can both control and interpret the experiments and their results properly (Guala, 2001, p. 461). Put generally, implementing experimental control means setting up the experiment so that no dis-

turbing factors affect the causal relationship being studied. Internal and external validity are two concepts used to interpret and analyze an experiment and its results (Jiménez-Buedo, 2011, p. 271-272). The concepts were originally coined by Cook and Campbell in the late 1950s, and have been under discussion since (Ibid, 272). Philosophers of science focus on analyzing the concept of external validity and its role in reasoning about causal claims and extrapolation in science. In economics, the focus of the external validity discussion is on assessing and developing methodology rather than conceptual analysis.

Roughly defined, internal validity describes whether the causal claims drawn on the basis of experimental results are reliable (Ibid, 271-272). It is a concept that describes whether the observed effects can be correctly attributed to their effects (Roe & Just, 2009, 1266). External validity, on the other hand, is a concept used to interpret experimental results as having the potential to be generalized to circumstances outside the experimental system (Jiménez-Buedo, 2011, p. 271-272). In this chapter, I present the discussion on external validity in philosophy of economics and consider the definitions given to the concept, as well as the arguments for and against its usage.

2.1 The Concepts of Validity

Put generally, validity in empirical economics describes whether a conclusion from an experiment is likely to be true, or whether it approximates the true conclusion or inference well (Roe & Just, 2009, p. 1266). The concepts of internal and external validity are used in disciplines from psychology to the social sciences, including behavioral and experimental economics[2] (Jiménez-Buedo, 2011, p. 272). In the

[2]The concepts of statistical conclusion validity and construct validity have been distinguished from internal and external validity (Jiménez-Buedo, 2011, p. 272). Statistical conclusion validity describes whether statistical methods are used correctly to determine relationship between vari-

methodological and philosophical literature, the definitions of and distinction between the concepts of internal and external validity are far from self-evident (Ibid, 272-273). In particular, it remains unclear what the different kinds of validity are attributes *of* (Ibid, 273-275).

Cook and Campbell, who developed the concepts, define internal validity as:

> "the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause" (Cook & Campbell, 1979, p. 37)

and external validity as:

> "the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings, and times." (Ibid.)

According to the definitions proposed by Cook and Campbell, internal validity describes whether the inference that the experimentally observed effect is causal is valid or not. It tells us, as Cartwright writes, "that in the experimental situation, the causal hypothesis is true" (Cartwright, 2007, p. 39). External validity describes whether the inference that the causal claim can be generalized to circumstances outside the experimental system is valid or not. This inference typically proceeds

---

ables. Construct validity refers to the validity of inferences from experimental particularities to general theoretical constructs. (Cook & Campbell, 1979, pp. 37-38) Roe and Just define ecological validity as the similarity of the context of an experimental task or situation to a presumed real-world choice situation (Roe & Just, 2009, p. 1267). Ecological validity can also be thought of as a dimension of external validity, along with other dimensions such as population validity and temporal validity – extrapolating to another population, or within the same population but to a different time (Guala, 2005, p. 142).

by comparing the model and target system for relevant differences and similarities (Guala, 2005, p. 180). If no relevant differences are found, the causal claim is considered externally valid. External validity can be interpreted as denoting a binary: either the experimentally observed causal effect remains or is likely to remain invariant in a specific target, across a range of targets or in general, or not (Deaton & Cartwright, 2018, p. 10).

Despite the attempts to define internal and external validity, especially the concept of external validity remains unclear. The concepts of validity, particularly external validity, are used to describe experiments, types of experiments, experimental results, experimental data, and experimental inferences (Jiménez-Buedo, 2011, p. 273-274). More recently, Marcellesi writes about "two properties of the conclusions", where the conclusions are drawn from experimental or nonexperimental studies and aim at estimating causal relationships (Marcellesi, 2015, p. 1308). Jiménez-Buedo concludes that the only way to use the two notions consistently is to use them to describe causal claims (Jiménez-Buedo, 2011, pp. 274-275). Experimental results are the observations and data gathered during an experiment, and causal claims concern the causal processes, mechanisms or dependencies whose existence or operation we infer from the experimental results (Guala, 2005, pp. 41-44). As such, causal claims (here taken as synonymous with causal inferences), not experimental results, are internally or externally valid. This is also the sense in which I will use the concepts of internal and external validity.

2.2 The Many Uses of External Validity

The two concepts of validity did not play a major part in discussions of experimental methodology until the late 1980s and early 1990s (Heukelom, 2011, p. 18-20). Vernon Smith, the key figure in founding and developing experimental economics,

wrote about precepts, or conditions, by which an experimental system is made into a real microeconomic system (Bardsley et al., 2010, p. 199). Experimental economists in the 1970s did not think that any division between the experimental system and a world "outside it" was accurate or methodologically useful. An experiment encapsulates the economic phenomenon itself, and inferences about the experiment hold for all economic phenomena of the same ilk. (Heukelom, 2011, p. 19-20). One of the precepts, parallelism, expresses this idea (Bardsley et al., 2010, p. 199-200). By the 1990s, the internal and external validity were used with regularity as contrasting concepts that experimental economists could use to "think about their experiments." (Heukelom, 2011, p. 22)

In the extant literature, analysis of external validity ranges from using the concept to evaluate the accuracy of predictions about the effectiveness of a policy program to detecting the factors that affect the generalizability and transportability of experimental inferences. In econometrics and political science, the concept of external validity surfaces in the study of causal inference, randomized experiments, and predicting the effectiveness of policy programs. External validity is quantitatively measured to evaluate the accuracy of causal predictions, and statistical methods for estimating external validity are developed. For example, Meager (2019) develops methods for quantifying the heterogeneity of the effects that compose the average effect, or average impact, of randomized controlled trials (RCTs) testing the effectiveness of microcredit. If it is heterogeneous effects that make up the average, predicting the impact of microcredit in a new context on the basis of the average will be "*at best uncertain and at worst infeasible* " (Meager, 2019, p. 16). Quantifying the heterogeneity of local effects means quantifying the external validity of the average effect (Ibid). This helps understand the impact of the tested policies, as well as the accuracy of predictions about the effectiveness of those policies in future sites.

In experimental economics, on the other hand, external validity is used more as a general term for identifying the factors, conditions and parameters that affect the generalizability or transportability of experimental inferences to "other circumstances" in general or a single target in particular. Still, the use of the concept of external validity varies also within experimental economics (Nagatsu & Favereau, unpublished, p. 20-22). Esther Duflo, one of the most notable figures of the evidence-based policy movement, and her coauthors define external validity in much the same way as Campbell: "*whether the impact we measure would carry to other samples or populations. In other words, whether the results are generalizable and replicable*"[3] (Nagatsu & Favereau, unpublished, p. 21; Duflo et al., 2007, p. 3950). Experimental economists not working with evidence-based policy focus on external validity partly because of the criticism aimed at laboratory experiments (Nagatsu and Favereau, unpublished). They want to include the "context-richness" that Loewenstein (1999) called for (Heukelom, 2011, pp. 21-22).

Referring to concerns of external validity, Loewenstein criticized experimental economists for a variety of reasons, one of which was that experimental economists tend to minimize the real-world content in their experiments, which then reduces external validity (Ibid, 22). Field experiments conducted by the "lab-minded" experimental economists answer these criticisms and aim to increase external validity by adding real-world context to the experimental design and setting. They also study whether laboratory results predict field results and vice versa. Analysis of external validity focuses on pinpointing the ways in which the phenomenon of interest the validity of claims about it are affected by changes in the experimental design (see e.g. Harrison & List (2004); List (2007); Levitt & List (2007a,b); Nagatsu & Favereau (unpublished).

[3] As the quote alludes, replicability and external validity are two intertwined concepts. Due to restrictions of space, I will not discuss their relation here.

2.3 Against External Validity?

Validity asks "how well the scientific operation achieves what it aims to achieve" (Heukelom, 2011, p. 13). Correspondingly, economists and philosophers treat internal and external validity as conceptual tools used to evaluate experiments (Jiménez-Buedo, 2011, p. 272). Some philosophers take a critical stance toward this and have critiqued the use of the concept in analyzing experiments. The criticism can be categorized into two intersecting strands. The first strand consists of criticism against the idea of external validity as a concept of generalizability or transportability. The second consists of criticism at the idea that external validity is a relevant concept for evidential reasoning, for example regarding experiments aiming at policy guidance.

Jiménez-Buedo (2011), and more recently Deaton & Cartwright (2018) and Reiss (2018) all argue that the use of the concepts of internal and external validity as a concept used to evaluate experiments should not be encouraged. Jiménez-Buedo argues that the concept of external validity is not a good concept of generalizability or transportability. It is vague, ultimately indistinguishable from internal validity in a meaningful way, and irrelevant for many kinds of experiments and their evaluation. (Jiménez-Buedo, 2011) Both Jiménez-Buedo and Westreich et al. (2018) argue that the division into internal and external validity is epistemically and methodologically unhelpful (Jiménez-Buedo, 2011; Westreich et al., 2018).

Other arguments fall into the latter category. Jiménez-Buedo uses ultimatum games as an example to argue that the majority of experiments in behavioral economics aim at testing competing explanations for observations of decision-making and behavior, for example altruism or prosocial preferences. Their point of behavioral experiments is not to yield empirical generalizations: the "most pressing questions" in behavioral economics do not "*pertain to matters related to the interference of confounds nor to the possibility of generalizing the findings to other settings ex-*

cept for in very general, mostly conceptual, terms." (Jiménez-Buedo, 2011, p. 279) Reiss argues that using the concepts of external validity encourages bad evidential reasoning, and proposes an alternative account of reasoning about target systems (Reiss, 2018, p. 9-18). External validity encourages bad evidential reasoning due to the concept's non-contextualist nature, and argues for a contextualist account of inferring causal claims about target systems. Like the other authors, Reiss points out the roles that experiments may have in scientific reasoning other than providing generalizable claims (Ibid).

Deaton and Cartwright argue that external validity does not necessarily matter for all experiments and use randomized controlled trials as an example (Deaton & Cartwright, 2018, p. 10-18). They argue that not all randomized controlled trials aim at externally valid results, and that experiments that fail to produce externally valid results should not be considered failures (Deaton & Cartwright, 2018, p. 10-18). Guala and Mittone point that exhibits are also experiments that not concerned with external validity (Guala & Mittone, 2005, pp. 510-511). "Exhibit" is a concept introduced by Robert Sugden, who writes that there are two kinds of experiments in behavioral economics: exhibits and theory-testing experiments (Sugden, 2005, p. 291). An exhibit is "*an experimental design which reliably induces some specific regularity (or "effect", or "phenomenon") in human behavior*" (Ibid). According to Guala and Mittone, preference reversals, ultimatum games, and dictator games fall into this category (Guala & Mittone, 2005). Exhibits are often not immediately concerned with issues of external validity or extrapolation, because they showcase anomalous behavior that is not explained by any received theory about decision-making and behavior (Sugden, 2005, p. 291).

The examples point to similar arguments. On one hand, a more contextual understanding of external validity would serve methodological analysis of experiments better. On the other hand, the concept is too vague to provide any relevant

insight regarding extrapolation. Furthermroe, the concept is not a useful concept in economics or its subfields, because many experiments are not concerned with generalization at all. The criticisms are correct in pointing out that external validity is a vague concept, both by definition and in use. They also point out problems in using external validity as the only concept to use when thinking about generalizability or transportability. It is also true that experiments have many roles in scientific inquiry, and that experiments aiming at formulating new hypotheses, testing theory, or showcasing anomalies may not be directly concerned with generalizations or external validity.

However, stating that behavioral economics is not concerned with empirical generalizations seems like a narrow view of the concept of external validity, or a narrow view of behavioral economics. If external validity is understood as a link between an experimental finding or inference and a domain where it could travel, then experiments that aim at capturing real-world phenomena correctly are, at least indirectly, interested in it. Linking an exhibit that showcases a behavioral anomaly to an empirically observed phenomenon, or use an exhibit as an "explanatory device" for patterns of behavior outside the experimental system (cf. Sugden, 2005, p. 298), requires explaining why the behavior showcased by exhibit (or some other behavioral construct tracked by behavioral economics and considered valid) is relevant to what is happening in the circumstance of interest. This involves comparing the experimental context in which the exhibit is observed to a real-world system – in other words, inferring external validity.

Similarly, Gneezy and Imas (2017) argue that experiments whose point is to collect facts to inform economic theory are also concerned with generalization (Gneezy & Imas, 2017, p. 441). If experimental data is used to test theoretical models and further develop them, then the experiment is concerned with issues of generalizability and transportability. The authors point to economic models of financial decision-

making, which were developed to represent how finance professionals, individuals investing for retirement, and other market participants behave. The experiments that tested these models in the laboratory used "a convenient sample" of undergraduate students, implicitly assuming that the students' behavior in the laboratory would be representative of and generalize to "the relevant population of experienced traders and financial market participants" (Ibid). The ways in which different kinds of experiments are interested in extrapolation thus remains an important question, despite criticism.

One way to gain more insight into the usefulness of external validity is to investigate its relationship with extrapolation. Guala's work on external validity highlights this relationship. He states that economists use the notion of "parallelism" to describe inferences from a specified model to a specified target (Guala, 1999, p. 569-570). He distinguishes the notion of parallelism from the notion of external validity, and argues that external validity is the notion of extrapolating inferences from inside a laboratory to another, unspecified system (Guala, 1999, p. 569). Later, Guala argues that according to the external validity hypothesis, the causal relata in a model system and the target system "belong to similar causal mechanisms" (Guala, 2005, p. 197). This is clarified by his reconstruction of the process of inferring external validity:

> 1. If all directly observable features of the target and the experimental system are similar in structure;
>
> 2. If all the indirectly observable features have been adequately controlled in the laboratory;
>
> 3. If there is no reason to believe that they differ in the target system;
>
> 4. And if the outcome of the systems at work (the data) is similar;
>
> 5. Then, the experimental and target systems are likely to belong to

> structurally similar mechanisms (or data-generating processes). (Guala, 2005, p. 180)

The problem of external validity as understood by Guala is thus not only about inferring or evaluating the superficial similarity of the model and target systems, but also about whether or not the data-generating processes in the target system are accurately captured in the model system. If the model and target systems are structurally similar with regard to relevant causal mechanisms, inferences from one can be generalized or transported to the other. Determining the level of similarity of the causal structures in the systems determines the inferences' generalizability and transportability. As chapter three will show, this is very similar to Steel's approach to solving problem of extrapolation.

Westreich et al. (2018) point out that the potential of an inference to be extrapolated is abuot "a relationship between a study sample and a target population for a particular question – rather than a single inherent characteristic of a study" (Westreich et al., 2018, p. 440). A target is always necessary for making claims of external validity, because the external validity of causal claims can change with regard to different targets (Ibid, 439-440). Without one, claims of external validity are not meaningful (Ibid, 440). Like Guala, Westreich et al. argue that when an experimental sample is nonrepresentative of the target population, then the external validity of causal claims is the outcome of an inferential process that compares the model to the target and defines the level of similarity between the two systems (Ibid).

If the potential to be extrapolated denoted by external validity is understood as a relation between a model and a target, the concept of external validity is relevant to evidential reasoning in experimental and behavioral economics. If an experiment is interested in producing causal claims that are externally valid with respect to some

target, then we can use the concept of external validity to evaluate the relationship between the experimental system and the target, and to what extent assumptions about similarities between the model and target systems are justified. In short, the concept has its use in evaluating the strengths and weaknesses of the experimental system as a model for real-world causal relationships with respect to particular targets. This does not assume that external validity is the most important concept in assessing experiments.

As such, the concepts are useful in identifying potential sources of error and bias in the experimental process (cf. Gerber et al., 2014, pp. 21-22). The argument is similar to Guala's (2012): worries about external validity are "*inescapable and indeed useful when addressed to the specific details of an experimental design, for in such cases they help establish the reliability of specific inferences from the laboratory to field settings*" (Guala, 2012, p. 7). External validity is a useful concept not only when discussing the potential effects of experimental design on whether inferences travel from the laboratory to a specific target, but also when establishing whether inferences travel from the lab to the field or one field setting to another.

In sum, whether an experiment is concerned with external validity, generalization, and extrapolation, depend in part on what purposes the experiment is conducted for, as well as the empirical phenomena the experiment is supposed to capture or represent and the experimental results illustrate. Rather than thinking about external validity as a non-contextual characteristic inherent to causal claims, the concept is better thought of as describing a context-dependent relationship between a model and a target system, specifically denoting the potential of causal claims to be extrapolated from one to the other. This relationship is not only about the superficial similarity between two systems, but whether the model system accurately captures the causal processes or mechanisms of interest in the target system. In general, experimental economics uses the concept to evaluate whether experimental

outcomes can be used for good explanations or predictions of economic phenomena.

Overall, the above discussion sheds new light on the idea that there is a single "problem of external validity". Instead, there are various problems of external validity regarding what it is, what it is useful for, and when. In economics and philosophy of economics, external validity it is a useful concept for analyzing and controlling the possibility of error. However, understanding extrapolation is key in understanding the kinds of epistemic risks the concept of external validity is used to mitigate. I turn to extrapolation in the next chapter.

3 Extrapolation and Field Experiments

3.1 Extrapolation as Comparative Process Tracing

Philosophy of social science usually takes the problems of external validity and extrapolation as synonymous. In this thesis, I will not do so. In philosophy of social science, the discussion has focused on conceptual and methodological analysis with the goal of constructing accounts of extrapolation that solve some of the epistemological and methodological questions related to generalizing or transporting causal claims from one context to another. The aim is to study, complement and inform science and scientific practice. (cf. Cartwright, 2011, 2012; Cartwright & Hardie, 2012; Guala, 1999, 2003, 2005, 2010; Jiménez-Buedo, 2011; Marcellesi, 2015; Khosrowi, 2019; Steel, 2008, 2010; Marcellesi, 2015; Reiss, 2018)[4]. Some argue that the

[4]Cartwright and Hardie's 2012 book *Evidence-Based Policy: A Practical Guide to Doing it Better*, like Cartwright's recent work on the topic in general, studies the epistemological and methodological questions related to making good effectiveness predictions in policy, but it does so from the viewpoints of a theory of evidence. It is thus mentioned as one of the works addressing problems of external validity and extrapolation, but will not be studied at length in this thesis, which focuses more on analyzing extrapolation as it appears in science also outside policy

problems of external validity and extrapolation are solved, others argue that they are not (Jiménez-Buedo, 2011; Cartwright & Hardie, 2012; Marcellesi, 2015; Reiss, 2018; Khosrowi, 2019). A review of external validity shows that it is a useful concept for assessing the generalizability and transportability of causal claims, but too vague to do all the theoretical weightlifting related to extrapolation. Now that I have discussed external validity, I turn to extrapolation itself.

Extrapolation asks what can be learned about a phenomenon within a target system on the basis of knowing something about the phenomenon within a model system (Steel, 2008, p. 78). Extrapolations are systematically used in science, also as epistemic tools to gain more evidence for or against a hypothesis (Baetu, 2015, p. 960-961). Theoretical accounts of extrapolation, such as comparative process tracing (Steel, 2008), analogical reasoning (Guala, 2005, 2010; Steel, 2010), and the more formal approach constructed by Pearl and Bareinboim (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2013; Pearl & Bareinboim, 2014) all aim at showing how extrapolation can work reliably. At their core, all strategies involve comparing the model system and the target system with analogical reasoning (Steel, 2010, p. 1058). The basic idea is that if the systems are alike, for example in their causal structure, then the causal effects will also be alike; the studied causal dependencies between given variables will remain invariant in the target. In comparative process tracing, the justification for drawing inductive conclusions between the model and the target is grounded in understanding the causal mechanisms in both (Steel, 2008).

In the ideal case, comparative process tracing yields information about significant causally relevant differences in the model and the target by tracing and comparing the nodes of the causal mechanisms in each. Mechanisms are compared to see where significant differences in their outcome are likely to occur, in order to assess whether the causal dependency stays invariant in the target. (Steel, 2008,

predictions and evidence-based policy.

pp. 87-92). Differences in the upstream nodes affect the downstream nodes, so only the downstream nodes need to be compared to yield information about the causally relevant differences between the model and target (Ibid, 79, 90). Comparative process tracing account differs from a naive mechanistic account of extrapolation, which argues only that information about mechanisms provides knowledge about how a cause produces its effect and thus warrants extrapolation (Ibid, 79).

Comparative process tracing is a fruitful approach to understanding extrapolation because it explains how information about mechanisms is information about causal structures, and how mechanistic information can be used to evaluate the relevant similarities and differences between causal systems for purposes of extrapolation. Both biology and social science, the fields Steel is interested in, are often interested in understanding the mechanisms underlying causal phenomena. The role of mechanistic knowledge in causal inference and extrapolation has been discussed in philosophy of social science (see e.g. Steel, 2004; Hedström & Ylikoski, 2010; P. K. Ylikoski, 2017; Marchionni, 2017; Marchionni & Reijula, 2019). The Tamil Nadu Integrated Nutrition Program (TINP) is used as an example of extrapolation that failed but would have succeeded had there been more knowledge about the social mechanisms in the target population (Cartwright, 2012; Cartwright & Hardie, 2012; Marchionni & Reijula, 2019). The goal of TINP was to reduce malnutrition in children by educating Indian mothers about children's nutritional needs. The implementation of a similar program in Bangladesh failed because there, it is the mother-in-law and not the mother who is in charge of feeding the child (Marchionni & Reijula, 2019, p. 56).[5]

[5]Steel formalizes his analysis of extrapolation as comparative process tracing using causal graphs, including directed acyclic graphs (DAGs). The use of DAGS and causal diagrams in extrapolation is also argued for by Marchionni & Reijula (2019). However, understanding them is not essential to understanding comparative process tracing, so I will not cover the use of causal graphs here.

3.1.1 Model and Target Systems

In the extant literature on extrapolation and external validity, few consider the exact roles that the target has in reliable generalization. Most use the concepts of "model syste" and "target system" as shorthands for "where inferences are extrapolated from" and "where inferences are extrapolated to". Reiss (2018) calls an experimental or a laboratory system about which something is inferred a model system. He defines "target system(s) of interest" as "Another, related system or set of systems – often a field system or a population different from the test population". This use of the concepts corresponds to for example Jiménez-Buedo's (2011) use of the concept of "target system".

If the model system is a laboratory system, an experimental setup or an observational study, it typically consists of a study design with which the phenomenon of interest, for example education, poverty or vote-buying, is investigated. Constructing an experimental design includes choosing an experiment type, the participants, the treatments, and randomization, among other things. In field experiments, the experimental participants are usually a subset of a larger population, for example a sample of the voters within a given electorate[6].

The concept of "target" can be abstract and undefined. For example, as mentioned, Guala argues that external validity is the applicability of causal claims, drawn on the basis of experimental results, to an unspecified set of target systems (Guala, 1999, p. 596). The notion of "target" is also used to mean a defined set of circumstances at a certain point in time. The target can also be thought of as "the real-world system (or set of systems) whose behavior we ultimately intend to investigate and understand" (Guala, 2005, p. 9). In economics, this is often a

[6]The size of the sample and its representativeness are much-discussed questions in experimental methodology and causal inference, but they are not always the only things that determine whether inferences about one population can be extrapolated to others.

"nonlaboratory entity", or an economy that is too big or complex to study directly (Ibid).

In this thesis, the phrase "model system" is used in a general sense, to refer to any system that is a "resembling representative" of a target system. This can be an experimental system or an observational study. (cf. Bardsley et al., 2010, p. 199). The phrases "target" and "target system" refer to the specific population, setting or other context to which causal claims are extrapolated to, unless otherwise specified. A more detailed, critical analysis on the concept of the target and its role in extrapolation is something that will, hopefully, be a subject of discussion in future research.

3.1.2 Challenges to Existing Accounts of Extrapolation

Steel builds his case for comparative process tracing on two epistemological challenges to extrapolation which an account of extrapolation should solve. The first is the *problem of heterogeneity*, or the fact that populations are likely to differ with respect to each other in causally relevant respects. The second is the *extrapolator's circle*, which asks for a way to get from inferences about the model system to inferences about the target system in a meaningful, non-circular way.[7] Steel argues that his account of extrapolation as comparative process tracing solves both. (Steel, 2008, p. 4, 7-8 79).

Solving the problem of heterogeneity means explaining how extrapolating inferences about one system to another is possible when there are causally relevant differences between the systems. The problem of heterogeneity is especially pertinent when experimental or observational studies are used to inform or design policy, because in general, policy reforms target or concern a certain population or sub-

[7]Both challenges were originally argued for as critiques against the methodologies of extrapolating from animals to humans, by LaFollette and Shanks (Steel, 2008, p. 4).

population. Testing those reforms, however, may be happen on a subset of these populations, other populations entirely, or with methods that hide the heterogeneity of effects within the population. If policy are tested on unrepresentative populations or in unrepresentative systems, there are likely to be causally relevant differences between the model and the target. An experiment or observational study that aims at providing results applicable for policy has to deal with causally relevant heterogeneity between, but also within, populations and subpopulations. An account of extrapolation has to explain how inferences from experimental and observational studies can be extrapolated even when the populations' observable and unobservable characteristics, the distribution of causal factors within those populations, and the heterogeneity of causal effects within and between the populations differ.

The extrapolator's circle is raised by the fact that if we knew all the similarities and differences between the model and target, conducting an experiment in a model system to investigate a target system would be useless, and extrapolation redundant. Warranting extrapolation requires knowing the relevant differences between the model and target systems, but extrapolation is meaningful only when we do not know what works in the target, i.e. whether there are relevant differences or not, and how those differences might affect the causal effect. Solving the extrapolator's circle amounts to justifying the use of a population as a model for the target population about which not much is known. (Steel, 2008, p. 4, 78)

Simple induction, mechanistic extrapolation, extrapolation based on causal powers and capacities, economic engineering, and analogical reasoning have been suggested as approaches to extrapolation (Steel, 2008, pp. 78-87). The general idea in all is similar to Guala's reconstruction of the process of inferring external validity: If the causal structures in the two systems are analogical to a certain extent, extrapolation from one to another is justified. Each account of extrapolation attempts at explaining what exactly is the most reliable basis for this analogy and its evaluation.

Should it be causal mechanisms, the causal capacities, or something else?

The most straightforward strategy for extrapolation is simple induction, which states that causal claims can be extrapolated to other circumstances unless there is a particular reason to think that something in the target is very different from the model (Steel, 2008, p. 78). Simple induction is an unsophisticated form of extrapolation, where the analogy between two populations relies on an unspecified criterion of "relatedness" (Ibid). Forms of relatedness, such as phylogeny or similarity of economic systems, are often not a sufficient warrant for extrapolation, so simple induction often yields mistaken extrapolations (Ibid, 80-82). As Steel shows, however, in cases where tracing social mechanisms is not possible, information about mechanisms can be necessary for conscientous and justified simple induction (Steel, 2008, p. 165-168).

Another way to address extrapolation is the notion of causal powers or capacities (Steel, 2008, p. 82). Think of a brick, which, by virtue of its physical properties, has the capacity to break a glass window. Cartwright (1994) argues that only knowledge of causal capacities allows us to extrapolate causal effects between populations (Steel, 2008, p. 82). Capacities are stable across background conditions, insensitive to variation in background variables. They include physical attributes and causal dependencies that are tied to complex sets of interactions, such as the capacity of aspirin to relieve pain. (Ibid, 82-83). Statements about capacities tell us what happens when the influence of all other causal factors is removed, but also more, because the capacity continues to influence the effect when there are other causes present, too (Ibid, p. 82). Even though a property "carries its capacities from situation to situation" (Cartwright, 1994, p. 146), Steel points out that the necessity for extrapolation arises when the stability of a causal relationship is doubted. In the end, the capacities account of extrapolation does not overcome the limits of simple induction. (Steel, 2008, pp. 82-85)

Economic engineering is the most reliable method of solving questions of extrapolation, because it turns the target system into the model system. The case of economic engineering most often mentioned in the literature in philosophy of economics is the Federal Communications Commission spectrum auctions, where economists were faced with the task of building an auction mechanism for allocating spectrum licences (Guala, 2001). Economists constructed an auction mechanism based on theory and tested it with experiments and simulations before implementing it (Ibid). Similarly, researchers at the California Institute of Technology designed and tested mobile phone auctions in the laboratory before the laboratory was exported to the real world. (Guala, 2003, p. 1204). If this strategy is followed, new institutions are created according to what works best in the laboratory (Reiss, 2018, p. 8). Reiss notes that this approach works best when the goal is to create new institutions and not explain existing ones, and when similarity between the new institution and the experiment is credible. (Ibid.)

Steel (2010) develops a generalized account of extrapolation, which he calls analogical reasoning. Analogical reasoning uses chain graphs, which are a tool for graphical representation that consist of nodes, lines and arrows (Steel, 2010, pp. 1062-1063). Chain graphs can be used for a variety of purposes, and Steel uses them to represent analogical inferences (Ibid, 1063). The nodes represent what is inferred, arrows represent causality and the lines represent analogy. For two things to be analogous, they have to be mutually similar (Ibid). Chain graphs can be used to illustrate this kind of similarity by drawing a line between two nodes, where the analogy is not mediated by any other node. For example, the chain graph $A \leftarrow B$ – $C \rightarrow D$ tells that B and C are analogical, so A and D are mutually informative. Borrowing Steel's example: I know Tom and Fred have similar temperaments. If I learn that Tom is prone to emotional outbursts, then I also learn it is probable Fred is prone to them as well. (Ibid) If I know B and C are analogical, learning A also

tells me D is likely.

By generalizing comparative process tracing into the account of extrapolation as analogical reasoning, Steel aims to also generalize his account of the evidence that is needed to support extrapolation by analogy. He refines the epistemological account of extrapolation that Steel (2008) began, but keeps the methodological discussion minimal. Extrapolation with information about causal mechanisms is still a pertinent question to the social sciences, as are the methodological issues of extrapolation it raises and their implications for the account of extrapolation itself. Accordingly, the rest of the thesis focuses on comparative process tracing and its applicability to the social sciences.

Comparative process tracing is based on the argument that causal mechanisms are the relevant causal structure whose understanding can help us extrapolate causal claims from one population or circumstance to another. It is a developed form of mechanistic extrapolation. "Naive" mechanistic extrapolation argues that mechanistic knowledge and knowledge of the factors capable of affecting the mechanism form a justified ground for extrapolation (Steel, 2008, p. 85). However, only mentioning mechanisms and assuming that they are analogical in the model and the target does not solve the extrapolator's circle, because we do not know how to yield information about mechanisms in the target without extrapolation becoming redundant. Neither does it explain how extrapolation is possible when there exist causally relevant differences in the relevant mechanisms. (Steel, 2008, p. 79, 85).

Guala (2010) reconstructs the three steps of extrapolation as comparative process tracing as follows:

> 1) Learn the mechanism in the model organism [or experimental system], by means of process tracing or other experimental means.
>
> 2) Compare stages of the mechanism in which the two [the experimental

> and the target systems] are most likely to differ significantly.
>
> 3) In general, the greater the similarity of configuration and behavior of entities involved in the mechanism at these key stages, the stronger the basis for extrapolation. (Guala, 2010, 1072-1073)

Steel understands process tracing as a method of inferring the causal mechanism between an input variable and the causal outcome or effect. This can proceed if we have observed a phenomenon being responsible for an outcome, for example HIV exposure causing AIDs. When we know the phenomenon as well as some constraints on defining the mechanism components and their interactions, we can infer the precise mechanism through which the phenomenon is responsible for the outcome by tracing forward from a point that is known as the starting point of the mechanism, or backwards from an end point, or both at once (Steel 2008, 87). In Steel's words, comparative process tracing solves the inference problem of the form: "*[g]iven both the mechanism and the phenomenon in the model, and partial information concerning the mechanism in the target, infer the mechanism and/or phenomenon in the target*" (Ibid).

Because of this, it is important to include Steel's point about antecedent background knowledge of the phenomenon in the target. The "central theme" of comparative process tracing is that background knowledge of the points of the mechanism where causally relevant differences are likely to arise is necessary for extrapolation (Steel, 2008, pp. 151-152). It is how the extrapolator's circle is solved. Only the use of relevant background knowledge explains how limited and partial information about mechanisms or phenomena in the target can be used for extrapolation.

The list of potential differences between populations is infinite, but ideally, all the causally relevant differences are captured in the causal structure, namely the mechanism. We can focus the comparison on the downstream nodes, because it

is there that the causally relevant differences between the model and the target populations will show. The extent to which the model must imitate the target and minimize the number of causally relevant differences depends on the specificity of the causal claim being extrapolated (Ibid, 8). In cases where we want to extrapolate general claims of causal relevance, a "total absence of causally relevant disanalogies" is not necessary (Ibid). This helps mitigate the problem of heterogeneity.

3.1.3 Mechanisms as Causal Structure

Steel understands causation in the interventionist sense, as invariance under an ideal intervention (Steel, 2008, p.11). Interventions are intentional manipulations of phenomena or systems (Ibid, 12). An ideal intervention implements a method of controlled variation that is unrelated to the studied causal dependency and leaves other causal relationships in the studied causal system intact (Ibid, 13). Steel himself argues that a manipulationist view of causation grasps well the way the biology and social science understand causation, but this does not make it the only useful account of causation (Ibid, 16).

The focus in comparative process tracing is on extrapolating inferences of positive or negative causal relevance. In addition to precise, quantitative claims about causal effects, such as those about the estimated effect of interest rates on inflation or years of education on income, we often want to know simply whether a cause is positively or negatively relevant to its effect (Ibid, 11, 19). Extrapolating claims of causal relevance can be formulated as, "*We know that X is a positive causal factor in the population P, and we want to know whether it is such in the distinct population P'* " (Ibid, 11). If we can establish with an intervention that changes in the values of X yield changes in the value of Y, the former is a causal factor of the latter.

This links causal relevance to probability: causal structures can be identified as

those that create probability distributions within populations (Steel, 2008, p. 37). According to Steel, it is "very compelling" to argue that if there is a causal dependency between X and Y, then changing the value of X will change the probability of Y (Ibid, 15). Thus, causal structures can also be used used to predict how an intervention affects probability distributions (Ibid). We get the following definition:

> (CS) Causal structure is that which generates probability distributions and indicates how these distributions will change given interventions. (Steel, 2008, p. 38)

Causal structure can be ascribed two roles: it generates probability distributions, and shows how interventions affect them (Ibid, 31).[8] Steel argues that a causal structure generating a probability distribution has to "exhibit behavior possessing the combination of individual disorder and aggregate regularity". This is one of the central notions in the foundations of comparative process tracing. Any phenomenon where aggregate regularities emerge out of individual irregularities can be represented in probabilistic terms and equated with generating a probability distribution (Ibid, 198-199). Mechanisms both generate probability distributions and indicate how the probability distributions change in response to interventions, so they are a plausible candidate for the relevant causal structure to base comparative process tracing on (Ibid, 32).

This does not yet sufficiently link probability to mechanisms, or illustrate how interventions on the causal structure (i.e. mechanisms) affect probability distributions (Ibid, 198). The disruption principle, another central notion in comparative process tracing, ties mechanisms as causal structure to the probabilistic concepts of causal effect and causal relevance (Ibid, 54, 198-199). It entails that if one can

[8] Probability is here understood as physical probability, not as degrees of belief (Steel, 2008, p. 38).

detect a relationship between X and Y by observing a change in the probability distribution of the latter given an intervention on the former, there exists at least one causal mechanism between the two (Ibid, 59). If there is a population within which X and Y are causally dependent, but within that population a subpopulation where they are not related, then the mechanism from X to Y is blocked within that subpopulation (Ibid, 54). The disruption principle explains how disrupted mechanisms do not produce probability distributions, and how undisrupted mechanisms do. Thus, Steel argues, mechanisms can be identified with causal structure especially in biology, and to a certain extent, the social sciences (Ibid, 31).[9]

Modularity is one of the conditions for causal structure, because when mechanisms are modular, information about mechanisms can also be used to understand the effects that interventions bring about (Ibid). If a mechanism is modular, then an intervention on one of the components of a mechanism leaves the causal generalizations about the other components unchanged (Ibid, 42). In other words, I can intervene on the mechanism from the light switch to a lit living room lamp by taking out the light bulb, but the other causal dependencies within the system, as well the causal generalizations governing those dependencies, remain invariant. Because modularity allows for the other causal relationships of a mechanism to remain unaltered, it helps predict the effects that interventions bring about (Ibid, 43-44).

Steel states that mechanisms are "*generally understood as consisting of interacting components that generate a causal regularity between some specified beginning and end points*" (Ibid, 40). According to Peter Machamer, Lindley Darden and Carl Craver, mechanisms are "*entities and activities organized such that they are pro-*

[9]The disruption principle follows from the principle of the common cause (PCC) and the faithfulness condition (FCC) (Steel, 2008, p. 55). Steel concludes that in general, assuming the PCC and the FC is not problematic. For further analysis of the plausibility of the PCC and the FCC in the social sciences as well as exceptions to them, see (Ibid, pp. 55-68).

ductive of regular changes from start or set-up to finish or termination conditions" (Machamer et al., 2000, p. 3).[10] Steel adopts this definition of mechanism, as it is the least insistent on any particular notion about causation and laws and how the two are related (Steel, 2008, p. 42). He focuses on regularly operating mechanisms, instead of singular or unique chains of events. (Steel, 2008, pp. 40).

Importantly, comparative process tracing is not only about comparing the similarity of causal mechanisms in two systems. The reason why Steel argues that extrapolation can be grounded in comparing mechanisms is that knowing the difference-making nodes of the causal mechanisms in the model and the target gives us information about the invariance of causal dependencies in the target. Comparative process tracing is not about comparing causal structure only to assess the similarities or differences between the causal structure in the model and the target, but doing so in order to determine the extent to which a causal dependency remains invariant in both. One critical question is, then, whether the similarity of causal structure is the relevant point of comparison when yielding information about the invariance of causal effects in different targets, and whether evidence about mechanisms is useful for conclusions about invariance. Due to limitations of space, these question will not be a central point of focus in this thesis. It is a question for further and future research, already studied in part by Marchionni & Reijula (2019).

3.2 Field Experiments in Economics

3.2.1 The History and Methodology of Field Experiments

In this section, I review field experiments and the particular issues of external validity and extrapolation to them. Field experiments are used for a wide variety of

[10] What mechanisms are and how they can be defined is has been and continues to be a central question in philosophy of science. Illari & Williamson (2012) provide one overview.

purposes and exhibit variety in terms of methodology (Gerber & Green, 2012, pp. 8-13). Field experiments range from small-scale behavioral games to large social experiments interested in policy recommendations. Some field experiments are interested in quantitatively evaluating the effect that microcredit programs have on poverty and others in testing how reciprocal voters in a Paraguayan electorate are, and whether their reciprocity is linked to politicians' behavior regarding vote-buying. The common element to all field experiments is that they adhere to standard experimental procedure, in which ideally only an intervention or treatment induces the operation of the studied causal mechanism or process, so that the effects of the causes can be accurately observed. This way, the experimenter can correctly interpret that the effects were only due to the treatment and not confounders. Field experiments differ from laboratory experiments by implementing elements of "fieldness", or "real-world context" in their design (Harrison & List, 2004; Levitt & List, 2007b, 2009). This one of the central things the discussion on the external validity and extrapolation of causal claims drawn on the basis of field experiments focuses on.

At its crudest, conducting a field experiment means conducting an experiment in which the setting that in some way resembles the actual setting where the studied phenomenon occurs (Gerber & Green, 2012, p. 10-11). This can be the electorate to which individual voters belong or a district where students live. However, the setting is only one dimension or criteria according to which the fieldness of an experiment can be evaluated (Ibid). Other dimensions include the resemblance that the treatment has to the intervention that would be applied outside the experiment; whether the participants are representative of the group that typically encounters the intervention; whether the context of the treatment is like the context of interest; and whether the outcomes of the experiment are similar to outcomes that would be of interest to either theory or practice (Ibid).

In short, most of the tangible elements of an experimental procedure (participants, intervention, setting) can be more or less field-like. In addition, the general context in which the experimental participants make their decisions can be field-like, as well as the task that they complete. The characteristics listed in the above quote can be summarized into four dimensions of fieldness: authenticity of treatments, participants, contexts and outcome measures (Gerber & Green, 2012, p. 11). Another way to evaluate fieldness is according to six criteria: the nature of the subject pool, the information the subjects bring to the task, the commodity, the task or trading rules applied, the nature of the stakes, and the nature of the experimental environment (Harrison & List, 2004, p. 1010).

The motivations for implementing elements of fieldness are varied. The field is not only better for studying certain policy-relevant phenomena than a laboratory, but studying phenomena within the field instead of a laboratory can also be used to infer the conditions under which different experimental types can be used as reliable tools of scientific inquiry (Harrison & List, 2004; Levitt & List, 2007a,b, 2009). Levitt & List (2009) identify the first generation of field experiments as beginning in the 1920's and 1930's with the work of Neyman and Fisher. They developed experimental methodology in general and applied it to field experiments in particular (Levitt & List, 2009, p. 2-4).[11]

The second generation of economic field experiments began with large-scale social programs implemented during and after the 1950's (Ibid, 2, 4-7). Levitt and List note that there are multiple definitions for social experiments, and draw on

[11]In addition to the generation-based categorization, a useful categorization is made by Nagatsu & Favereau (unpublished), who argue that the development of field experiments has happened in two simultaneously developing strands. The first consists of experimentalists interested in policy, and who now conduct randomized field experiments and policy evaluations. The second consists of experimentalists who conduct field experiments after moving to the field from the laboratory. (Nagatsu & Favereau, unpublished, pp. 4-14)

previous definitions to present social programs as experiments that include elements of control, policy intervention and statistical analysis, and aim at drawing causal claims about the effects of a change in policy (Ibid, 4). Accordingly, the early social experiments measured either structural parameters or typical behavioral relationships in areas like employment, national health insurance, and social benefits, and were used to evaluate the effects of public policies related to these. (Ibid, 6.) As examples of early social programs, Levitt and List mention a study on the effects of electricity pricing in Britain from 1966 to 1972, and the New Jersey Income Maintenance experiment, whose purpose was to investigate the effects of negative income taxation (Ibid, 4-5).

Third-generation field experiments are smaller-scale experiments aiming to experiment on 'naturally-occurring populations in naturally-occurring settings' (Ibid, 7). Field experiments are now used for a variety of purposes: theory-testing, fact collection for theory construction and providing behavioral principles to sharpen inferences from the laboratory or help with the interpretation of laboratory results or uncontrolled data. In addition, they can be used to provide data helpful for analyzing the causes or underlying conditions that produce the experimentally observed phenomenon (Ibid). Different institutional settings, such as schools, police precincts, public housing projects and voting wards all serve as fertile ground for field experiments (Gerber et al., 2014, p. 10).

Harrison & List (2004) compile a taxonomy of field experiments that is based on the level of control implemented by the experimenter. They divide field experiments into artefactual, framed and natural field experiments, where lab experiments are on one end of the scale, and natural experiments on the other. Field experiments, which include multiple different kinds of designs, fall in the middle. Artefactual field experiments most closely resembling lab experiments and natural field experiments resembling natural experiments. (Harrison & List, 2004, p. 1013) According to

some, field experiments provide a bridge between laboratory data and naturally occurring data, because they include both experimental control and realism (see e.g. List, 2007). This mixture is usually not possible in either laboratory or natural experiments. (Levitt & List, 2009)

Artefactual field experiments are field experiments conducted much like a laboratory experiment, but with an experimental population that typically consists of participants from the context of interest (Harrison & List, 2004, pp. 1013-1014). For example, Henrich et al. (2001) conduct ultimatum, dictator and pubic goods games in small societies in developing countries to see whether the behavioral patterns are similar to those observed and replicated in industrialized countries and modern societies. The factors which the experimenter can control, such as payoffs and information given to the subjects, are held nearly identical in all fifteen societies. The results show that there are major differences between the participating societies, from which correspondence to differences in everyday life and social norms can be inferred. (Levitt & List, 2009, p. 8)

Framed field experiments are otherwise similar to artefactual field experiments, but they formulate the experimental tasks, traded goods, stakes and the information given to the subjects according to the setting where the studied phenomenon naturally occurs (Harrison & List, 2004, p. 1013-1014). The primary motivations of framed field experiments are theory testing and policy guidance (Levitt & List, 2009, p. 9). Natural field experiments are experiments that are conducted where the studied phenomenon naturally occurs (Harrison & List, 2004, p. 1013-1014). The experimenter has minimum or no control over the experimental design, as the idea is to observe controlled comparisons that occur naturally (Ibid, 1041). The subjects do not know that they are being observed and ideally, behave as normal. This, in the words of Levitt and List, "combines the most attractive elements of the lab and naturally-occurring data: randomization and realism" (Levitt & List, 2009, p. 9).

Some of the most well-known field experiments in economics today, including the framed field experiments conducted by Duflo and her coauthors, are randomized trials, often called randomized controlled trials, randomized field experiments, or policy evaluations. They have been used to study many kinds of phenomena, including child malnutrition, the effective distribution of anti-malaria bednets, the effect of microcredit on poverty, business training, and the impacts of educational programs among others. In principle, randomization mitigates the problem of confounders (the problem of identifying observable and unobservable factors that could potentially affect outcomes) because it ensures that the likelihood of the observable and unobservable factors being present in the groups is equal in both (Gerber & Green, 2012, p. 7-8).

Experimental designs can also be purposely combined and layered to see how different designs impact the study (Harrison & List, 2004, pp. 1009-1014; List, 2006; Nagatsu & Favereau, unpublished). An experimenter can conduct multiple experiments with differing experimental designs, and compare the outcomes to infer that the changes in them are due to changes in the experimental design. If some part of the experimental process – the participants, the setting, the task – is changed to be more "field-like" and all else is kept equal, then any changes in experimental results should occur because of changes in field-likeness. This tells us that laboratory results might not generalize to the field or vice versa. (Ibid)

For example, the gift exchange experiment in List (2006) is a nested experiment with an artefactual field experiment, a framed field experiment and a natural experiment conducted to investigate gift exchange and prosocial preferences in the marketplace and compare results between the experiment types. The results show that the prosocial preferences, which emerge in both the artefactual field experiment and the framed field experiment, do not emerge in the natural experiment. (Levitt & List, 2009, p. 10). Another experiment investigates the effectiveness of

fundraising appeals using a laboratory experiment and a field experiment (Gerber & Green, 2012, p. 10). The correspondence between the results in the laboratory and the results in the field is "relatively weak", as the laboratory results do not predict the field results very well. (Ibid). All this has been central to the discussion of the advantages and disadvantages of field experiments, often phrased in terms of external validity.

3.2.2 Advantages and Disadvantages of Field Experiments

The main advantage of conducting field experiments is related to external validity. They are argued to provide results from which externally valid claims can be inferred. The idea is that claims about the processes or effects observed in a field experiment can more easily be extrapolated. However, the relationship between field experiments and extrapolation turns out to be more complex when different perspectives are brought into discussion with one another.

As field experimentation grew more popular, experimental economists started to contest the notion that laboratory experiments provide inferences that are inherently generalizable to non-laboratory settings (see e.g. Harrison & List (2004); Levitt & List (2007b). Laboratory experiments strip the phenomenon of interest of its context, but the context can be relevant for the observed behavior and its interpretation in many ways (Ibid). Field experiments study the target system itself, so the results are generalizable to the target system (Bardsley et al., 2010, p. 243). On the other hand, the set of experimental participants is rarely truly representative of the general population they are drawn from, at least in randomized trials (Westreich et al., 2018, pp. 439-440). This also leaves the validity of transporting claims about the experimental results to external populations unwarranted (Bardsley et al., 2010, p. 243; Westreich et al., 2018, pp. 440). Cartwright and Hardie (2012) and Deaton and Cartwright (2018) also both point out the problems in arguing that randomized

controlled trials provide inherently generalizable outcomes and inferences.

Nevertheless, compared to laboratory experiments, field experiments are considered advantageous, and this is a central motivation for conducting them (Harrison & List, 2004; Levitt & List, 2007a,b, 2009; Gerber et al., 2014). This line of argumentation often includes reference to the artificiality of laboratory experiments. The argument is that laboratory experiments simplify, abstract and idealize the phenomenon of interest as well as its setting and the causal factors affecting it, and field experiments aim to represent a more realistic, complex and diverse environment by adding "real-world" context to the experiment (Gerber & Green, 2012, pp. 8-16; Gerber et al., 2014, p. 23; Harrison & List, 2004; Jimenez-Buedo & Miller, 2010; Levitt & List, 2009; Schram, 2005). Field evidence is called for because it is seen as an antidote to the problems laboratory experiments can have with external validity and extrapolation (e.g. Gerber et al., 2014, pp. 22-23; Guala, 2012) This has led to argumenst that laboratory experiments have high internal validity, but low external validity, and that for field experiments, the tradeoff is vice versa (cf. Jimenez-Buedo & Miller, 2010, p. 302, 313-314; Gerber et al., 2014, pp. 22-23). High internal validity in the laboratory comes at the expense of lower certainty about the laboratory depicting the properties of interest in the phenomenon of interest (Bardsley et al., 2010, p. 242).

On the other hand, Jimenez-Buedo & Guala (2016) problematize the notion of "artificiality" when it is applied to laboratory experiments, showing that the concept itself as well as its supposed effects on generalizability and transportability remain vague. Bardsley et al. also note that artificiality can refer to many things: "the isolating function of the lab, its potential contaminating effects, or its alteration of objects of investigation" (Bardsley et al., 2010, p. 214). Jimenez-Buedo & Miller (2010) problematize the idea of a tradeoff between internal and external validity. Using an experiment from behavioral experimental economics as an example,

Jiménez-Buedo and Miller conclude that replicating the experiment with slight variations to identify potential confounds remedies issues of both internal and external validity. Therefore, it does not seem logical that there necessarily exists a tradeoff between internal and external validity. (Jimenez-Buedo & Miller, 2010, pp. 317-319). Instead, internal validity issues precede questions of external validity, because if the inferences are not internally valid, there is no meaningful way to extrapolate them (Jimenez-Buedo & Miller, 2010, p. 319; Santos, 2011, pp. 49-50)

In addition, field experiments do not automatically remove the possibility of interaction effects, demand effects and other types of bias to affecting the experimental results that are associated with the artificiality of laboratory experiments. (Jimenez-Buedo & Guala, 2016, p. 19). Furthermore, Gneezy and Imas also point out that due to their naturalistic context, field experiments may be inherently situation specific, and thus results are harder to replicate in new settings. This may, in fact, contribute to difficulties in generalizing experimental outcomes or inferences to new systems, or in comparing them to other populations or settings (Gneezy & Imas, 2017, p. 440). Nonetheless, the nested experiments mentioned at the end of the last section support the idea that including or excluding field-like elements can impact the outcome of an experiment as well as the potential of the experimental inferences to extrapolate to new targets. In laboratory experiments, the treatment effect is easier to detect correctly, and the experiments are easier to replicate and it is easier to compare the outcomes to those of other populations (Gneezy & Imas, 2017, p. 440).

The advantages for using field experiments to increase external validity and ease extrapolation proves to be a cluster of interrelated methodological questions related to the purpose of the experiment and the possibilities of controlling error when reasoning inductively from experiments. To sum up, one could say that the issues any kind of experiment has with internal or external validity or extrapolation

depends on what is being studied, and what the target is. On one hand, this highlights the fact that experimental economists also treat inductive reasoning from experiments as an epistemic risk, because there is always uncertainty related to the accuracy of conclusions regarding potential targets. On the other hand, this highlights the relevance of understanding experimental methodology in different contexts to understanding problems of extrapolation. The next section looks at a particular methodological solution that aims at maximizing the benefits of both laboratory and field experiments by combining elements from both.

3.2.3 Lab-Like Field Experiments for Policy

One of the central motivations for conducting lab-like field experiments instead of field experiments in naturalistic settings is that they solve the tensions between artificiality and external validity (Viceisza, 2016; Gneezy & Imas, 2017). Lab-like field experiments include artefactual field experiments and framed field experiments, when the latter are conducted in a field context (Viceisza, 2016, pp. 835-836). Similarly, Gneezy and Imas define a lab-in-the-field study as "*one conducted in a naturalistic environment targeting the theoretically relevant population but using a standardized, validated lab paradigm*" (Gneezy & Imas, 2017, p. 440). Lab-like field experiments combine elements from field experiments and laboratory experiments. They maximize the benefits and minimize the costs of both types of experiments: They use participants from the field, which increases the potential of the experimental inferences to extrapolate, and they use a laboratory procedure, which enables control over causal factors and confounders. (Gneezy & Imas, 2017, p. 440)

There are a few reasons to study extrapolation from lab-like field experiments. They are often conducted with policy concerns in mind, especially in the context of development economics (Viceisza, 2016, p. 836). Secondly, in the current literature, the discussion about extrapolation in experimental economics has focused on

laboratory experiments on one side, and field experiments studying behavior in naturally occuring contexts on another (Ibid). A reason for analyzing field experiments with comparative process tracing is that lab-like field experiments can be used in conjunction with randomized controlled trials or other studies to yield information about the causal processes and mechanisms leading to observed behavioral effects (Viceisza, 2016, pp. 836, 842-843; Gneezy and Imas, 2017, p. 448). Collecting explanatory covariates can help identifying the hypothetical mechanism that drives the success or contributes to the failure of the planned intervention or policy program. Policy makers can then use data from both experiments to plan beneficial policy interventions. (Gneezy & Imas, 2017, p. 450-452) In other words, this information is useful for extrapolation.

Lab-like field experiments seem like an ideal experiment if a researcher is interested in outcomes about which generalizable or transportable causal claims can be drawn from and extrapolated. They are not only concerned with the discovery of causal mechanisms, but can also be used for that purpose. Information about causal mechanisms is useful for causal explanation, prediction, and the design and implementation of policy interventions (cf. Ibid, 450-452). Looking at lab-like field experiments with comparative process tracing can not only shed light on the applicability of comparative process tracing as an account of extrapolation in social science, but also on the specific methodological issues of extrapolation that field experiments interested in causal mechanisms face. The next section investigates whether comparative process tracing is applicable to field experiments in economics in general, and the fourth chapter studies lab-like field experiments in particular.

3.3 Field Experiments and Comparative Process Tracing

Can comparative process tracing explicate and be applied to field experiments in economics? AS mentioned, in principle, there is no reason why comparative process tracing should not apply to or yield useful results in the social sciences. However, applying comparative process tracing to actual cases reveals the limitations of its as an account of extrapolation regarding them. In this section, I first discuss why comparative process tracing is a fruitful approach to understanding extrapolation in the social sciences. Then, I review the challenges it faces.

Comparative process tracing is a useful account of extrapolation because it is meant to be a general account of mechanisms-based extrapolation, applicable across disciplines. It details when, why how extrapolation can be based on causal mechanisms information about causal mechanisms. As the case of TINP illustrates, this information is often valuable in itself, and also as a complement to other kinds of evidence. As discussed, field experiments can be used for a wide variety of purposes, including the investigation of causal mechanisms and processes underlying observed causal effects. Comparative process tracing thus seems like a fruitful approach to understanding extrapolation of causal claims about causal mechanisms observed in field experiments.

Even though comparative process tracing is meant to apply to both biology and in social science, Steel states that it is likely to not work in social science as well as in biology (Steel, 2008, p. 9) There are two reasons for this. First, in the social sciences, there is more uncertainty regarding whether an intervention will turn out structure-altering. Structure-altering interventions are interventions that cause non-modular changes in social mechanisms. If information about social mechanisms does not tell us the effects of an intervention, the issue may be that social mechanisms do not fulfill the conditions required of causal structure. (Ibid)

The second challenge is that there may be uncertainty about the causal mechanisms responsible for a given phenomenon of interest. This complicates extrapolating claims for one circumstance to another, because we may not know that the mechanism responsible for the causal dependency is the same in both (Ibid). For example, there are two possible mechanisms to explaining preference reversals. Whether one or the other is the correct one significantly changes our understanding of how widely preference reversals are spread "outside the laboratory walls" (Ibid). The applicability of the conclusions regarding the emergence of preference reversals in the real world depends on which mechanism is correct (Ibid, p. 9, 169-174).

In an example, Steel illustrates the practical difficulties of using comparative process tracing in social science. The case he considers is a welfare-to-work program, implemented in various states in the United States, and evaluated with a randomized controlled experiments. The goal of the studies was to guide changes in welfare policy on a national level, and they were used as empirical evidence when policymakers made changes in the federal welfare program (Steel, 2008, p. 163, 166). According to Steel, one reason why comparative process tracing does not explicate extrapolation in the case of the welfare program example is that randomized controlled trials, in general, are good for estimating causal effects but not necessarily for providing evidence about the social mechanisms producing those effects or the potential of claims about those effects to generalize. (Ibid, 163). In this case, however, it seems that comparative process tracing is being applied to the wrong kind of experiment. Comparative process tracing can be a useful account of extrapolation in an experimental context if the experiment provides evidence of social mechanisms.

Additionally, Steel doubts whether claims about the positive causal relevance of welfare program can be extrapolated to a national scale or to different economic circumstances, even with information about mechanisms. Mechanisms on larger scales are likely to differ, so any mechanistic information from the model might not

help us infer what will happen on larger scales. (Ibid, 165-166). He compares the welfare program experiments and the case of aflatoxin and argues that in the case of aflatoxin, one can study metabolism in rats and compare it to metabolism in humans, either with in vitro studies or blood samples (Steel, 2008, p. 166). It is not certain that information about the likely similarities and differences between the experimental system and a target ares similarly available in the case of the welfare program (Steel, 2008, p. 166).

Additionally, tracing the relevant stages of the mechanism might not be possible. (Ibid, 166) The problem is that you can only observe the operation of a program at the location of its implementation, and because of this, "it is unclear that comparative process tracing can facilitate extrapolation to new locations or larger scales", which is what one is concerned with in the case of welfare reform (Ibid). However, the issue of scalability is at the heart of the problem of heterogeneity in social science: randomized trials and smaller, non-randomized experiments are conducted to see how an intervention will work in a particular population or location, and to transfer that knowledge to other populations. It cannot be the reason that any particular account of extrapolation fails, for it is the reason why accounts of extrapolation are developed in the first place.

In addition, Guala argues that the lack of a concrete target is why extrapolating preference reversal mechanisms fails. Without first identifying a target, comparative process tracing "cannot even take off" (Guala, 2010, 1080). He suggests that instead of focusing on preference reversals in ideal competitive markets, they should be studied in real economic settings where they are likely to manifest (Ibid). However, as the last section showed, field evidence may not ameliorate the problem of extrapolation. I agree that identifying a target is necessary for comparative process tracing, but also because it helps specify the relevant background knowledge necessary for comparative process tracing. My argument is similar to that of Reiss (2018). If the

goal is to extrapolate the experimental findings to new contexts but a target is not specified, it is unclear what kind of evidence an experimenter should be looking for in the experiment in the first place, and with which methods.

Comparative process tracing begins by tracing the mechanism in the model system. This amounts to providing a mechanistic *explanation* about the phenomenon within the model. Kuorikoski and Ylikoski argue that only by knowing the explanatory task at hand can one determine which details of the mechanism are relevant for explaining the phenomenon, and the right abstraction level (P. Ylikoski & Kuorikoski, 2010, p. 52). Mechanism-based explanations seek to capture the necessary elements of a causal process and abstract away irrelevant details. If something does not make a difference to the causal effect being explained, it can be left out. (Hedström & Ylikoski, 2010, 50).

When explicating the cogs and wheels of a causal mechanism for purposes of extrapolation, knowing the target beforehand is useful because the phenomenon in the target is what we are interested in explaining or making predictions about. By specifying a concrete target, the extrapolator has more access to relevant background information regarding the context in which the causal effects should stay invariant. In the paradigmatic biological example of aflatoxin B1 causing cancer in rats, the question is whether the animal model can reliably be used for extrapolating the claim that aflatoxin also causes cancer in humans. Knowing the target helps choose a suitable model and a suitable method for investigating social mechanisms, if extrapolation with mechanisms is what we are interested in.

There is also a conceptual argument to be made. As discussed, in comparative process tracing, mechanisms are the relevant causal structure to compare between two systems insofar they generate probability distributions and provide information about the changes that an intervention brings about. For comparative process tracing to work as an approach to understanding extrapolation, mechanisms and their

nodes have to be traceable on the basis of experimental evidence and relevant background information. Steel uses the Machamer-Darden-Craver -definition of mechanism, in which mechanisms are "*entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions*" (Machamer et al., 2000, p. 3). He defines social mechanisms as, "*complexes of interacting individuals – usually classified into specific social categories – that produce regularities among macrolevel variables*" (Steel, 2008, p. 48). Marchionni (2017) modifies the definition and defines economic mechanisms as, "*complexes of rational agents, usually classified into social categories, whose actions and interactions generate causal relationships between aggregate-level variables*" (Marchionni, 2017, p. 425)

Problems in extrapolating social mechanisms might not be the result of the mechanisms not filling the conditions of causal structure, but of the concept of "mechanism" and "social mechanism" and their entailments being unclear. Kuorikoski (2009) introduces two concepts of mechanism that are used to discuss and model mechanisms in the life sciences and social sciences. These are mechanisms as componential causal systems (a CCS) and mechanisms as abstract forms of interaction (an AFI). Componential causal systems consist of "*a set of component parts fulfilling different causal roles within a nearly decomposable system*" (Kuorikoski, 2009, p. 147). Taken together, are responsible or realize a causal property in the system. The regularity and external validity exhibited by the system is due to the intrinsic causal powers of the mechanism components and the stability that results from their organization. According to Kuorikoski, the Machamer-Darden-Craver definition of mechanism is an example of a CCS. (Ibid, 147)

An AFI, on the other hand, is a mechanism that explains some macro-level property or phenomenon of interest, but one where the outcome cannot be as straightforwardly decomposed and its causal functions, localized into component

parts (Ibid, 150-154). Mainstream economics is often interested in understanding macro-behaviors made up of the micro-level behaviors of individuals, households or firms maximizing their utility (Ibid, 150). However, the macro-level phenomenon is the result of the "cumulative interactions and interdependencies of the parts", rather than a streamlined causal path (Ibid). The relevant causal properties of each micro-level unit, such as preferences, are not separable from the system the mechanism is embedded in (Ibid).

Nonetheless, AFIs are structures that consist of parts, their operation and their functions, which together produce a macrolevel phenomenon (Ibid, 151-152). They just cannot be *traced* in the same way as CCS mechanisms. Explaining the macro-level phenomenon by matching "component operations with different component parts [...] with varied intrinsic causal powers" is not possible. This may cause difficulties for process tracing (Ibid, 153). It follows that extrapolation as comparative process tracing would also face challenges when dealing with AFI mechanisms. An example Steel gives of a social mechanism is Schelling's segregation model; it is also what Kuorikoski uses as an example of an AFI mechanism (Steel, 2008, p. 49; Kuorikoski, 2009, p. 151).

In other words, one of the problems in transferring comparative process tracing from biology to social science could be that Steel's emprical analysis of causation is done with the Machamer-Darden-Craver -definition of mechanism in mind, but his notion of a social mechanism assumes different causal facts and entails different theoretical consequences about mechanisms and reasoning with them (cf. Kuorikoski, 2009, p. 143). The issues in using comparative process tracing as an account of extrapolation in the social sciences could also be within the mechanism concept, used to reason about the conditions of tracing mechanisms to reach conclusions of invariance, and not in the nature of social mechanisms.

4 Case Study

4.1 From Reciprocity to Respect for Earned Property

The case study investigates the methodological aspects of applying comparative process tracing to lab-like field experiments interested in social mechanisms. In philosophy of economics, the cases that the debate on extrapolating experimental causal claims has mainly focused on are auction experiments, most notably the aforementioned Federal Communications Commissions spectrum auctions (cf. Alexandrova, 2006; Guala, 2001, 2005, 2010). To recap, they were a case of economic engineering, where the auction mechanism was constructed on the basis of theoretical knowledge, tested with simulations and experiments, and then transported to the real world. The target was created as an analogical system to the model.

I wanted to conduct a case study in order to understand in more detail why comparative process tracing could or could not be applied to experiments in the social sciences. Because of the discussion on laboratory experiments and randomized field experiments, I wanted to look at other kinds of field experiments interested in a policy-relevant phenomena and social mechanisms. Examples of fitting experiments were provided by two review papers, Viceisza (2016) and Gneezy & Imas (2017). I went through the experiments listed in the review papers and chose two studies that used lab-like field experiments to study social mechanisms responsible for the phenomenon of interest. Both examples in the case study use laboratory and field evidence to yield conclusions about the phenomena of interest and the social mechanisms underlying them; a strategy which Guala (2012) calls for. Scrutinizing the studies through a lens of extrapolation and comparative process tracing illustrates the usefulness of mechanistic information for extrapolation, but also the challenges that comparative process tracing faces in practical instances of extrapolation in economics.

The first study is noted by Viceizsa (2016) as a study that opens the black box of causality, or in other words, discusses the possible the mechanisms through which an observed effect operates. The study provides evidence that the reciprocity of individual voters helps sustain vote-buying by affecting who politicians target when buying votes. In addition, it suggests mechanisms that explain why and how politicians target reciprocal individuals. The study draws inferences about the way vote-buying is sustained by voters' reciprocity. Understanding the mechanisms that sustain vote-buying is beneficial because if the mechanisms are uncovered, then the practice can be monitored and its effects on electoral discipline weakened (Finan & Schechter, 2012, pp. 863-864). The results are of interest to the phenomenon of vote-buying in general, not just the vote-buying happening in the study setting.

The second case investigates a study where a randomized policy evaluation tracking academic achievement is supplemented by a lab-like field experiment. It is mentioned by Gneezy and Imas (2017) in their chapter on lab-in-the-field methodology as a case where a lab-like field experiment is used to complement a randomized policy evaluation in order to shed more light on the processes mediating the effects observed in the randomized study. The authors conclude that the experiments provide evidence that academic achievement causally affect individuals' values. In this case, the specific changes are observed individuals' respect for private property rights.

It is important to note that even though the experiments rely and draw from behavioral economic theory, they are not experiments meant to lend support or falsify either standard economic theory or any behavioral economic theory per se. Both experiments investigate social preferences, which have been studied in behavioral economics from a variety of perspectives. Social preferences are "*other-regarding preferences in the sense that individuals who exhibit them behave as if they value the payoff of relevant reference agents positively or negatively.*" (Fehr & Fischbacher,

2005, p. 151). The aim of the studies is to investigate reciprocity and respect for property rights, and how they are related to the macrolevel phenomena of vote-buying and education.

4.2 Vote-Hungry Politicians Target Reciprocal Individuals

4.2.1 The Study and Its Central Outcomes

Finan and Schechter define vote-buying as the act of a voter exchanging their vote for material goods or other forms of redistribution (Finan & Schechter, 2012, p. 864). Understanding the interpersonal relationships between voters and politicians is crucial when studying vote-buying because the process happens through personal interaction (Ibid). The authors argue that vote-buying can be maintained by an internalized norm of reciprocity, because receiving money or favors from politicians can foster feelings of obligation (Finan & Schechter, 2012, 863). They combine survey data with evidence from a lab-like field experiment to show that politicians, presumably knowing this, target reciprocal individuals, which provide evidence in support of their conclusion: Individual reciprocity can help sustain vote-buying. (Ibid, 864)

The authors use a cooperation game to measure the average reciprocity of the members of the electorate in which vote-buying is known to happen. Reciprocity in vote-buying is modeled as a problem of cooperation. The problem of cooperation has a long history in political philosophy and philosophy of social science (Guala, 2012, p. 2-3). Economists and biologists studying reciprocity posit that there exist two different mechanisms: strong reciprocity and weak reciprocity (Ibid, 1-2). According to weak reciprocity models, choosing reciprocal strategies must profit the agent who plays them. In contrast, according to strong reciprocity models, the strategy that a strong reciprocator plays does not have to be profitable. Strong reciprocators can

cooperate in a game even when it would be more profitable to exploit the other players. They can also punish players who choose not to cooperate at a cost to themselves. Guala calls these two types positive strong reciprocity and negative strong reciprocity respectively. (Ibid.)

Finan and Schechter define reciprocity as either intrinsic or instrumental. Intrinsic reciprocity is "*a person's willingness to sacrifice his own material well-being to increase the payoffs of someone who has been kind to him, or to decrease the payoffs of someone who has been unkind to him*" (Finan & Schechter, 2012, p. 864). This corresponds to strong reciprocity (Finan & Schechter, 2012, p. 864, fn. 6; Sobel, 2005, p. 397, fn. 3) Instrumental reciprocity is understood as reciprocity that is motivated by forward-looking self interest, and this corresponds to weak reciprocity (Ibid). It is intrinsic reciprocity that causes "a kind act by one individual" to "affect the preferences of another to elicit kindness in response" (Finan & Schechter, 2012, p. 864, fn. 6). According to the authors, their experiments measure intrinsic reciprocity rather than instrumental reciprocity. One-shot and anonymous play, where the game is played only once and players are kept unknown from each other, removes the strategic considerations linked to weak reciprocity (cf. Ibid, 869).

The central finding is that middlemen, the people employed by politicians to offer individuals goods in exchange for their vote, are likely to target reciprocal individuals rather than nonreciprocal ones. The result does not change when the effect of other factors are controlled for, and neither is intrinsic reciprocity a proxy through which other cooperation-sustaining mechanisms operate. Instead, the experiment measures a feature of the voter's utility function, and this feature is more likely to make the voter reciprocate noninstrumentally. The measurement of reciprocity is "strongly and robustly" correlated with targeted vote-buying. (Ibid, 864)[12]

[12]The authors state, "*A 1 standard deviation increase in reciprocity increases the likelihood of experiencing vote-buying by 44%. This finding is robust to controlling for a rich set of individual*

The authors suggest three plausible explanations for why politicians target reciprocal individuals, thereby opening the "black box". The first explanation is that vote-buying solves issues of commitment in voting. Standard models of elections often suggest that vote-buying should not exist because secret ballots ensure that votes remain unobserved and a politician's votes unenforceable. There is no formal way of contracting votes, so vote-buying by targeting reciprocal individuals is the solution for problems raised by anonymous voting. The second explanation is that politicians know the voters' preferred party, and are paying them to actually vote. Politicians have an incentive to target reciprocal individuals because they can be paid less than their disutility from voting (as receiving money creates a want to reciprocate), whereas for a non-reciprocal individual, the price is equivalent to their disutility. (Ibid, 877) There are two possibilities: Middlemen would either be dissuaded from targeting individuals with a high propensity to vote, because they will vote independent of being offered goods, or individuals with a low propensity to vote, because buying their vote will be more expensive. The authors cannot test for either because they do not have a measure of someone's propensity to vote. (Ibid, 877-878)

The third explanation is that the voters view voting as a repeated game, in which reciprocity helps sustain cooperation (Ibid, 878). This could happen if participants did not think their votes were secret; nonetheless, middlemen have an incentive to target intrinsically reciprocal adults (Ibid). Importantly, if a voter thinks that the ballot is not secret, then they may respond to kindness with instrumental (weak) reciprocity. They might sacrifice their immediate payoffs, and be reciprocal towards a middleman, "*not because he is intrinsically reciprocal, but to sustain a long-term relationship with the middleman or his party.*" However, the authors state

characteristics, including other social preferences as well as social network architecture." (Ibid, 864)

that they can find no evidence of the experimental results reflecting instrumental reciprocity (Ibid, 878-879).

4.2.2 Method and Experimental Procedure

The study uses data from an experiment and from the field, and is able to link the reciprocity the authors measure experimentally to actual political behavior (Ibid, 866). The authors gather survey information on vote-buying as experienced by members of the electorate in the 2006 municipal election, and combine it with experimental data on voters' individual instrinsic reciprocity (Finan & Schechter, 2012, p. 864-869). They use data from a household survey conducted in March-July 2007, which was the fifth round of a study that began in 1991. The 1991 survey selected 300 households from 15 villages in rural Paraguay. (Ibid, 867) Over time, some of the participants left the study. In 2002, 223 households remained, and 187 of these sent a member to participate in experiments that measured risk aversion, trust, and trustworthiness. In 2007 new households as well as new study components to capture voting and vote-buying were added. In total, 140 individuals both participated in the experimental and political components (Ibid, 868).

This way, the authors are able to link participants' experimentally measured reciprocity to information on vote-buying in the 2006 elections in the cases of 140 players, out of the 187 original participants. (Ibid, 868-869) The authors also conducted surveys with people who act as middlemen between politicians and voters in order to grasp whether middlemen knew the voters well and whether the political operatives offered individuals goods in exchange for their vote. The study has two measures of vote-buying: one from data from the household survey and one from the interviews with middlemen, which were conducted in 2010. (Ibid, 868)

The experimental component of the study is a trust game with two players

(Ibid). As mentioned, all games were one-shot and anonymous (Ibid, 869). In the game, the first player was given 8000 Gs and they had to decide how much of it to send to the second player. The alternatives were to send nothing, 2000, 4000, 6000, or 8000 Gs. However much the first player sent was tripled, and the second player could either keep all or return a sum of her choice. Before finding out the sum that was sent to her, the second player had to choose how much she would return if she received 6000, 12000, 18000 or 24000 Gs, and play as she had chosen. (Ibid, 868)

The second player chose how much she wants to return according to her level of altruism and to her level of reciprocity. Finan and Schechter explain: "*The more altruistic they are, the more they should return in all four cases. The more reciprocal they are, the more they should return when the first mover treats them well and the less they should return when the first mover treats them poorly.*" (Ibid, 868-869). If a player is altruistic, then they return more money in all four situations. If the player is reciprocal, they send money back reciprocating the first transaction: the less money is sent, the less is returned.

The authors assume that the second player thinks she has been treated well if first player sends at least half of her Gs. If the first player sends less, then the second player thinks that she has been treated poorly. (Ibid) The authors calculate a measure of reciprocity by calculating the average share of the total sum returned when the second player receives 12000, 18000 or 24000 Gs (in other words, when they receive half or more of the money available to her). Then, the share that the second player returns when they have received 6000Gs (in which case the first player sent only a quarter of their money) is subtracted from the first average. This means that altruism is subtracted out and a pure measure of reciprocity is achieved. (Ibid, 869)

4.2.3 Challenges to Comparative Process Tracing

Evidence from the lab and the field leads the researchers to conclude that sustained vote-buying is a result of social mechanisms. Following Marchionni's definition, there are complexes of rational agents, classified into social categories (voters, middlemen and politicians), whose actions and interactions create relationships between aggregate-level variables (average reciprocity and vote-buying). Accordingly, Finan and Schechter propose specific mechanisms through which politicians and middlemen can target reciprocal individuals and which sustain vote-buying. At first glance, this information is relevant and helpful if one wished to extrapolate claims about reciprocity upholding vote-buying to other contexts.

Before discussing extrapolation to new contexts, it is worth to note that the study already relies on extrapolation. The authors assume that claims about the outcomes of the lab-like field experiments hold for the whole target population. Finan and Schechter choose to implement a lab-like field experiment likely because they can use a sample of the relevant population, keep the experimental procedure tightly controlled, and combine the results with data from the field. Questions of generalizing from the experiment to the field, related to sample size, the representativeness of the experimental task, and other factors, are partly answered by the choice of experimental design and the use of statistical methods.

Guala (2012) discusses some of the problems in assuming that reciprocity mechanisms are generalizable from laboratory to field settings. According to him, models of negative strong reciprocity (i.e. models of costly punishment for free-riders) have a wider scope of application, but it is not clear that they accurately capture mechanisms of reciprocal behavior outside the laboratory, or just experimental artefacts. Field data does not support inferences about behavioral patterns of negative strong reciprocity drawn on the basis of laboratory data, so the experimental outcomes

regarding the phenomenon are likely artefactual. (Guala, 2012) Artefactual does not mean "not real", but that the experimental outcome occurs as a result of the experimental conditions, and not as a result of correctly manipulated causal mechanisms that also exist outside the lab. (Ibid, 6-7) Models of weak reciprocity (i.e. self-interested strategic cooperation) are better corroborated by empirical evidence than models of negative strong reciprocity, but their scope of applicability is narrower. Models of weak reciprocity apply to situations where the future over which the player strategizes is long, there are only few players, and the flow of information in the group is unrestricted. (Guala 2012, 2, 3-4, 13-14).

Guala also introduces the notions of a "narrow" and "wide" reading of reciprocity experiments. According to the uncontroversial narrow reading, punishment experiments are tools used to measure social preferences (or "robust psychological propensities") (Guala 2012, 2, 5-6). According to the wide reading, punishment experiments capture the mechanisms that support cooperation also in the real world, outside the laboratory. It implies that costly punishment mechanisms are what sustain cooperation in real-life situations. The two readings of reciprocity experiments should be kept separate, for the mechanisms that sustain cooperation outside the laboratory may be very different than mechanisms that sustain cooperation outside it (Ibid).

The experiment conducted by Finan and Schechter is not an experiment of negative strong reciprocity, but the outcomes support claims of positive strong reciprocity. On one hand, Finan and Schechter treat their experimental outcomes in accordance with the narrow reading. They use the experiments to measure villagers' reciprocity. The experiments are used as methodological tools, to turn "unobservable attitudes and dispositions ("preferences") into observable and quantifiable experimental variables". On the other hand, while not interested in studying the true nature of positive strong reciprocity, the authors are interested in linking experi-

mentally observed reciprocity to reciprocity in real political life. To this end, the authors do precisely what Guala recommends: They find data from other sources to corroborate the experimental findings, and link behavior exhibited by the participants during the experiment to actual political behavior. The authors are able to argue that that the consistency of the measure of reciprocity with field data supports the conclusion that individual voters' reciprocity helps to sustain vote-buying. Because of this, the relevant question is rather whether the measure of reciprocity is accurate, and claims about measured reciprocity internally valid.

What about extrapolating claims about the reasons why politicians target reciprocal individuals to new circumstances? The authors suggest that mechanistic evidence is relevant for understanding and explaining the existence of vote buying. As previous discussion has shown, it can also be relevant for extrapolation. However, anyone wishing to extrapolate claims about the mechanisms between reciprocity and vote-buying faces the practical challenges mentioned by Steel (2008) and Guala (2010): There are many possible mechanisms responsible for the phenomenon, and no concrete target is identified by the authors themselves. The authors suggest three mechanisms responsible for the relationship between reciprocity and vote-buying. Vote-buying is a solution to commitment issues, it is politicians knowing the voters' party preferences and paying the voters to turn up to vote, or it helps sustain cooperation in a repeated game. Any correct mechanistic explanation about the relationship is underdetermined by the available evidence, because the evidence supports all explanations. Depending on the target, extrapolating the wrong claim can have unintended consequences. In general, this might not be a problem, because knowing that politicians target reciprocal individuals already gives some information about possible interventions should someone wish to end vote-buying in their own district. In Steel's terms, we know that reciprocity is relevant for vote-buying. If the turnout model is correct, then this might be a bigger problem, because the

available evidence does not tell whether it is voters with a high propensity or a low propensity to vote that the politicians, or anyone wishing to end vote-buying, should target.

The study does show how knowing a target for extrapolation is necessary for comparative process tracing to take off. If no target has been identified, knowledge of the causally relevant similarities or differences is harder to identify. As mentioned before, Henrich et al. (2001) show that the outcomes of behavioral games can vary in different communities, even when the experimental procedure is held nearly identical every time. The difference in results corresponds to differences in everyday life and social norms (Levitt & List, 2009, p. 8). In addition, Jakiela (2011) shows that the outcomes of behavioral experiments can vary substantially depending on the cultural context they are conducted in. A target would indicate some of the conditions under which the effect of reciprocity on vote-buying should remain invariant, for example the size of the community, the magnitude of vote-buying, and the structure of the process through which politicians and political parties buy votes.[13] Additionally, the study illustrates the many roles experiments can have in social science. Rather than provide directly extrapolatable claims, the aim of the study is to provide new information about a policy-relevant phenomenon, and show that the results can be used to construct testable hypotheses about potential mechanisms that operate between voters' reciprocity and politicians targeting them. Its outcomes are policy relevant precisely in the way some of the critics of external validity point out.

Overall, the case study provides a more detailed illustration of the importance of a target for extrapolation, in order to determine the relevant background knowledge

[13]It is important to note that I argue that information about the target can be used to collect and organize relevant information about the context, following of Steel emphasizing the role of background knowledge information. I am not advocating for deciding the target in order to know causal mechanisms in it, so the extrapolator's circle is not triggered.

necessary for comparative process tracing. In this case, knowing the target is useful also for determining the consequences of extrapolating uncertain claims about social mechanisms. On the whole, the study suggests that understanding the problems of extrapolation that stem from experimental methodology, related to internal validity as much as transportability, also help understand the questions related to the epistemic justification of extrapolation. For example, questions of structure-altering interventions are not relevant to the study, because it does not propose claims about interventions which one could extrapolate. However, interventions planned on the knowledge that reciprocity is a relevant factor for vote-buying could run into problems of epistemic justification, if we plan a large-scale intervention on a small-scale experiment. In sum, understanding the methodological limitations of different kinds of experiments regarding extrapolation also helps identifying the knowledge we have access to and need in order justify extrapolation.

4.3 Academic Achievement Affects Social Preferences

4.3.1 The Study and Its Central Outcomes

Jakiela et al. (2015) use a lab-like field experiment to measure the effects of a policy intervention, a scholarship competition implemented in a random sample of primary schools in Western Kenya, on the social preferences of the individuals targeted by it. The program's effect was evaluated by the non-governmental organization that implemented it, and it was shown that participation in the program led to improvements in standardized academic tests (Ibid, 390-391). In order to measure the effects of academic achievement on social preferences, the authors conduct a lab-like field experiment where participants from the treatment and control groups play a variant of the dictator game. The authors use an instrumental variables method[14] to iden-

[14]The instrumental variables method is a method of causal inference from statistical data. To

tify the causal impact of participating in the program on social preferences (Ibid, 386-387). The authors use a dictator game to test whether girls who participated in the educational intervention are less likely to appropriate income earned by others for themselves. This acts as a measure for changes in social preferences, namely respect for property rights. (Ibid, 386-388)

The central finding is that people from the treatment group of the educational policy are less likely to appropriate another player's earned income, or in other words, that education affects social preferences (Ibid, 388). The finding is relevant for both, educational and economic policy (Jakiela et al., 2015, p. 386) Results from the laboratory game show that there are clear differences between participants in the treatment group with more education attained, and the participants in the control group, with less education attained (Ibid, 396). The participants in the former allocate more to the player whose income is being divided than the participants in the latter. The changes brought about by education cannot be explained by changes in beliefs or in generalized altruism, but by changes in social preferences as measured by changes in earned property rights (Ibid, 388,) The authors conclude that the causal relationship between the educational intervention and behavior is mediated by a mechanism of academic achievement (Ibid, 401-402).[15]

There are two explanations for how academic achievement could mediate this

---

infer causality between X and Y, a researcher uses a third variable, Z, that is a cause of X but not of Y, and an effect of neither. In addition, Z cannot share a common cause with either, and any effect of Z on Y passes through X. (Steel, 2008, pp. 178-179, 183) In this case, allocation to a treatment school is an instrumental variable with respect to social preferences, because it is independent of background factors that could by themselves impact social preferences (Jakiela et al., 2015, p. 387).

[15]The authors state, "*Point estimates suggest that a one standard deviation increase in academic test scores is associated with a 10 % point increase in the share of the budget allocated to other.*" (Ibid, 388)

effect that the authors themselves consider most plausible. First, increases in academic achievement (human capital) could alter respect for earned property rights (social preferences) by individuals learning to embrace certain values through operating in an educational environment where the exertion of effort is incentivized and rewarded, and the benefits from said effort are private (Ibid, 402). The other option is that an individual might observe that academic success is a signal for later success, and chooses self-serving moral codes because they believe that high productivity should be rewarded (Ibid).

4.3.2 Method and Experimental Procedure

Primary school in Kenya lasts for eight years, and ends in the Kenya Certificate of Primary Education exam (Jakiela et al., 2015, p. 389). A student has to successfully complete this exam to be admitted to secondary school, but nearly all students take the exam, whether they are planning on continuing school or not. The intervention whose effects were assessed by the policy evaluation was the Girls' Scholarship Program (GSP), an education initiative targeting girls in the grade standard 6. It was implemented in 34 primary schools in 2000. The program aimed at improving the academic performance of adolescent girls, and encouraging to stay them in school. (Ibid.) In 2001 and 2002, the ICS awarded scholarships to every girl standard 6 in the Busia District who scored in the top 15% of the KCPE practice exam (Ibid, 390).

The impacts of the education initiative were evaluated in a randomized program evaluation. 69 primary schools were randomly assigned to the treatment group or the control group, and the schools in the latter did not participate in the initiative. In 2001, 110 girls received scholarships, and a year after, the program was implemented for the second cohort of girls. Citing Kremer et al. (2009), the authors state that the program improved academic performance in the KCPE practice exams among girls

but also boys, and at all performance levels. The spillover effects may be explained by increases in teacher attendance. (Jakiela et al., 2015, p. 390)

The experiment that measures the effect of the policy intervention on social preferences is a variant of a dictator game (Ibid, 394). Dictator games measure how willing participants are to share, and they have been used to measure the beliefs, norms and ideals that ground conceptions of fairness in dividing income (Ibid, 387).[16] Dictator games are games where one player is the "dictator", who is allocated a certain amount of money. The dictator then decides how to divide the amount between themselves and another player. (Ibid.)

In the version conducted by the researchers, one of the participants earns money by completing a task, and this money is then divided between the dictator and the other by the dictator herself (Ibid, 394-395). The dictator's task is to appropriate their chosen sum *out of the income of the other player*. Using one's earned income instead of unearned income generates an informal property right, and the experiment measures changes in the respect for this property right. (Ibid, 387-388) Other previous dictator game experiments suggest that more educated participants are not more altruistic or generous, but do seem to have increased respect for others' earned property rights (Ibid, 395). Additionally, the authors combine the experimental data with data from various sources: administrative data tracking individual test scores prior to and during the intervention, and student surveys in both treatment and control schools, and a follow-up survey administered to all women from both the 2001 and 2002 cohorts from the treatment and control schools. (Ibid, 391).

[16]The trust game in the previous example is a dictator game with the sum decided by the player who makes the first transfer.

4.3.3 Challenges to Comparative Process Tracing

On the whole, the challenges that using comparative process tracing as an approach to understanding extrapolation are similar in both studies, due to their similar methods and aims. Finan and Schechter use a trust game to measure levels of strong reciprocity of individuals. They combine this evidence to other data to show that the individuals who behaved reciprocally are also the ones targeted in vote-buying. Jakiela et al., use a dictator game to measure individuals' respect for private property, and the results of this experiment are coupled with field data and used as evidence for the claim that social preferences were changed by the educational intervention. Again, the study provides evidence that there exist mechanisms through which academic achievement affects an individuals' respect for earned property rights. The mechanisms consist of complexes of rational agents, classified into social categories, whose actions and interactions create causal relationships between aggregate-level variables, namely educational attainment and willingness to appropriate another individuals' earned income. Again, this information is relevant for comparative process tracing.

In addition to the previous example, the study by Jakiela et al. depends on extrapolating from the experimental sample to the whole of the two age cohorts. Again, this is less of a question of mechanistic extrapolation than a question of the choice of experimental design, the use of the experiment, and the implementation of statistical methods. Extrapolating from the sample to the two age cohorts is about correctly measuring this change in social preferences. As the discussion in the previous section shows, it is more related internal validity than correctly inferring psychological mechanisms of respect for property rights and extrapolating them. Furthermore, questions of generalizability are partly answered, again, by the choice of experimental design. The authors conduct a lab-like field experiment because it means that the participants are drawn from the relevant population, but that the

experimental procedure can be strictly controlled. Like Finan and Schechter, Jakiela et al. use the experiments as tools for measuring social preferences. They are not meant to capture *why* dictators allocate something to the other player. Neither are the experiments meant to capture mechanisms responsible for the effect between academic achievement and social preferences.

Further discussion on using dictator games as measurement tools is provided by Guala & Mittone (2010). They discuss the instability of results from dictator games, which arise even when results from games with similar subject pools and countries with similar social structures are combined. Results from dictator games may not be robust, which casts doubt on the use of the games as a measure of respect for property rights (cf. Ibid). (Guala & Mittone, 2010, pp. 578-580). Dictator games yield volatile results because their design is so simple and abstract that they fail to elicit a specific norm, and the experimental participants fail to observe any norm that they should act in accordance with (Guala & Mittone, 2010, p. 581). There are two reasons why this is not likely a concern for the validity of the outcomes of the experiments conducted by Jakiela et al. First, the doubts cast on the use of the dictator game as a tool of research mainly concern dictator games that are meant to yield evidence that either supports or falsifies standard theory (cf. Guala & Mittone, 2010, p. 581). This is not the point of the experiments that Jakiela et al. conduct.

Second, dictator games may elicit specific norms when their design is altered (Guala & Mittone, 2010). For example, Cherry et al. (2002) add a "legitimacy factor" to the dictator's assets by making the dictator earn the money, instead of simply getting the money (Guala & Mittone, 2010, p. 597). This gives the participants reference to norms that they encounter in situations in real life (Ibid). The experimental design triggers a "powerful normative mechanism that invites people to behave in a self-interested manner" (Ibid, 581). This is comparable to the modifications to the dictator game design made by Jakiela et al., where the money to be

divided is not windfall money, but earned by the player not in the dictator's role.[17]

What about extrapolation to new contexts? Extrapolating claims from the study with comparative process tracing faces similar challenges as extrapolating claims from the vote-buying study. The randomized field trial and lab-like field experiment opened the black box of the educational policy, so the population from which the study sample of girls is drawn could act as one concrete target. Even then, the question remains: Is the relevant target for generalization here the two cohorts who participated in the intervention, or some bigger set of the population of the Busia district? Jakiela et al. note that a parallel randomized experiment in the Teso district, a neighbor of the Busia district, did not provide conclusive evidence regarding the success of the GSP scholarship competition. This was partly due to difficulties in program implementation. (Jakiela et al., 2015, p. 390) In light of the evidence available, the target population to which experimental inferences from the study could be generalized is restricted to the Busia district.

No concrete target external Busia is identified, so whether information about causally relevant similarities or differences between the study system and a target is available is not self-evident. If the target system where inferences about the effects of educational attainment are extrapolated to a district in Finland, different background information about contextual factors is needed than if the target system is another district in Kenya. Additionally, the available evidence supports many explanations about mechanisms between academic achievement and social preferences. It depends on the target whether extrapolating the "wrong" mechanism has any consequences, like Steel argues that extrapolating the wrong preference reversal mechanism has. Last, in the previous study, the results remain policy-relevant even though they are not directly transportable to any new circumstance. Overall, similarly to the study by Jakiela et al, the study illustrates in more detail how extrap-

[17]The design Jakiela et al. use is innovative, and has not yet replicated well.

olation is a question of both experimental methodology and epistemic justification. Understanding both is key in understanding how the problem of extrapolation can be solved in social science.

4.4 Comparative Process Tracing as Extrapolation in Economics

Steel's example of extrapolating claims about the carcinogenic effects of aflatoxin B1 illustrates how comparative process tracing works as an account of extrapolation in biology. The examples in this case study have shown that in experimental economics illustrate the methodological challenges that extrapolation, specifically extrapolation as comparative process tracing, faces in experimental economics. Applying comparative process tracing to examples of studies interested in social mechanisms shows that many challenges to extrapolation are about finding the right experimental methods to suit one's epistemic goals. Analyzing whether comparative process tracing can account for extrapolation in economics shows that understanding the practical challenges to extrapolation that stem from economic methodology can complement our understanding of the epistemological challenges to extrapolation in social science, and how they can be overcome.

On an argumentative level, the authors of both studies present their findings in a way that suggests that their findings are relevant to theoretical and empirical research in general, rather than just the site-specific experiments conducted in both studies. "*We argue that in rural Paraguay, vote-buying is sustained, in part, by intrinsic reciprocity [...] our findings provide evidence on the influence that reciprocity can have in politics*", Finan and Schechter argue (Finan & Schechter, 2012, p. 879). Jakiela et al. write, "*We provide evidence that increases in human capital, as captured in academic achievement tests, alter individual values, generating greater*

respect for earned property rights. This finding demonstrates that formal education can have cultural impacts..." (Jakiela et al., 2015, p. 404).

In principle, comparative process tracing is a very useful account of extrapolation in social science. It explains when and how information about causal mechanisms can be used to aid extrapolation. Social science is often interested in explaining social phenomena with the mechanisms that produce them, and using that information to extrapolate findings from one context to another seems ideal. For example, identifying the social mechanisms operating in a field experiment could increase knowledge of the invariance of causal effects across contexts and help extrapolation.

A conceptual challenge related to applying comparative process tracing to the social sciences concerns the definition of "mechanism" employed by Steel. The main takeaway from the studies is that the social mechanisms identified by both Finan and Schechter and Jakiela et al. are the relevant causal structure about which more information could be helpful for extrapolating claims about the observed effects to new contexts. The nature of social mechanisms might complicate transferring comparative process tracing to social science like Steel suspects, but an unclear grasp of the relevant concept of mechanism and consequently, the relevant mechanistic knowledge, might complicate it as well.

In addition, it is important to note how scientists themselves use the concept of mechanism. Marchionni (2017) introduces four ways in which the concept of mechanism is used in economics, originally explicated by Reiss (2013). First, econometricians use the notion of mechanisms to distinguish causality from correlation. Second, mechanisms are understood as the intervening variables between a cause and its effect. Third, mechanisms are understood as underlying structures or processes, for example the market. Fourth, mechanisms are understood as pieces of theory that explicate, for example, the conditions for some economic phenomena. (Marchionni,

2017, p. 423; Reiss, 2013, p. 104-105) Marchionni refers to the two latter notions as mechanisms as underlying structures, and states that they are the ones that come closest to how mechanisms are commonly understood by mechanistic philosophers now (Marchionni, 2017, p. 243). Philosophical discussions about causal inference and extrapolation do not always keep this understanding of mechanisms separate from the understanding of mechanisms as intervening variables (Ibid).

It is not clear that the concept of mechanism relevant for Steel is similarly relevant for experimental economists looking into social mechanisms, as it is for researchers interested in biological mechanisms. For example, in the studies mechanisms between reciprocity and vote-buying and educational attainment, academic achievement and social preferences are treated more like underlying structures than intervening variables. Reiterating Kuorikoski's argument about two kinds of mechanism concepts, it is not entirely clear that the knowledge about mechanisms that Steel states is necessary for extrapolation as comparative process tracing actually matches the mechanistic knowledge that is relevant for social scientists, or possible for them to gain. Consequently, it is not clear that comparative process tracing can account for what kind of information about mechanisms is necessary for extrapolation and concluding that a causal stays invariant in new targets.[18]

From a methodological standpoint, the examples illustrate the practical challenges stemming from experimental methodology that using comparative process tracing as extrapolation in experimental economics faces. Even when an experiment is interested in and able to explain a phenomenon with mechanisms, extrapolating claims about the mechanisms with comparative process tracing is not entirely possible. As mentioned, Steel himself suspects that comparative process tracing

[18]Nicholson (2012) discusses the different notions of mechanism in philosophy of biology. He discusses the disparities in philosophers' and biologists' understanding of mechanisms, and the implications those disparities have for philosophy of science interested in mechanisms.

may not work as well in social science as it does biology. The examples in the case study show that this skepticism is more a byproduct of the way he sets comparative process tracing up as an account of extrapolation, rather than the reasons Steel himself suggests. As it stands, comparative process tracing is an epistemologically comprehensive account of extrapolation, but it does not meet the practical side of experimental inquiry in economics. Scientific inquiry depends on extrapolation, so understanding the methodological and practical issues related to it enhances any theoretical or epistemological account on the topic.

The case study highlights the importance of the target population for mechanistic extrapolation. Conclusions about the usefulness of knowing the target are similar to those made by Reiss (2018). He argues for a pragmatist, contextual epistemology of evidential reasoning. Reasoning about causal dependencies within target systems should "begin with a hypothesis about that system, and ask what types of evidence we need to establish that hypothesis" (Reiss, 2018). Reiss also points out the different practical functions experiments may have in scientific inquiry: they suggest hypotheses, provide direct support for hypotheses and specify existing hypotheses (Reiss 2018). Cartwright and Deaton (2018) also voice similar concerns with regard to randomized controlled trials. They argue that randomized controlled trials can be used for a variety of purposes, and not all of those purposes include or require the application of experimental results to new targets (Deaton & Cartwright, 2018, p. 10). Randomized controlled trials, like all experiments, are ways to learn different kinds of things. They can provide counterexamples to general theoretical propositions, confirm predictions of theory, or used as evaluation procedures, to show stakeholders that a project achieved its goals (Ibid, 13). These functions are exemplified by the examples.

In addition, the case study illustrates that different types of experiments can be concerned with different types of extrapolation. Khosrowi (2019) distinguishes

between attributive extrapolation and predictive extrapolation, and argues that the latter is more prevalent in evidence-based policy. Attributive extrapolation attempts "attribute an observed effect causally to its suspected causes" (Khosrowi, 2019, p. 50). Predictive extrapolation, which is the kind of extrapolation usually encountered in evidence-based policy, aims to "predict the future effects of (interventions on) suspected causes" (Ibid). Put simply, attributive extrapolation is interested in conclusions about the causes of effects in a particular study setting, and predictive extrapolation aims at reaching predictive inferences about the effects of future interventions on a different target (Ibid). Predictive extrapolation is utilized in situations where the intervention of interest (such as an educational reform or a welfare reform) has not been implemented in the target, or its effects in the target not observed (Ibid).

Steel's welfare program example and the studies conducted by Finan and Schechter and Jakiela et al. illustrate the difference. Steel is not concerned with finding out the causes behind changes in well-being in a welfare program or welfare reform. Instead, he is concerned with extrapolating causal claims about the effects of an intervention to contexts where the intervention has yet to be observed. The extrapolation that the examples in the case study are concerned with, on the other hand, resembles attributive extrapolation. Both studies are concerned with studying the causes of effects in a particular study setting, namely the factors behind vote-buying in rural Paraguay, and the causes behind changes in academic achievement and social preferences in Kenya.[19]

I do not mean to argue that comparative process tracing fails as a mechanistic account of extrapolation, or that extrapolation based on information about

[19]Khosrowi argues that comparative process tracing does not work very well when predictive extrapolation is the aim, but because his arguments concern econometrics and evidence-based policy specifically, I will not discuss them here.

causal mechanisms is never relevant for economics or for policy; on the contrary. The case study does not diminish Steel's arguments regarding the importance and usefulness of mechanistic extrapolation for social science, but details the specific challenges mechanistic extrapolation faces where it could be used, namely policy-relevant studies interested in learning about causal mechanisms. The case studies show that a solution to "the problem of extrapolation" faces challenges on multiple levels. There are the general, epistemological puzzles that Steel scrutinizes, the extrapolator's circle and the problem of heterogeneity. On a related level, there are the challenges that specific methods such as field experiments, as well as the phenomena being studied, pose for identifying causation, mechanisms, and relevant background evidence for extrapolation.

Last, there are the questions related to the practical aims of different studies. Policy-relevant research are broad, one experiment is just an experiment, and experimental evidence one type of evidence among many. The case study shows that inferences about the outcomes of behavioral experiments may not be straightforwardly applicable to policy-making or immediately transportable to new contexts, but can nonetheless provide insights for policy. Results and inferences from experimental and observational studies can provide testable hypotheses for future studies. To conclude, comparative process tracing remains a useful epistemological account of extrapolation, but it needs to be complemented with a specific methodological account of how to address the kinds of challenges mechanistic extrapolation faces in practice, and with respect to different goals.

5 Conclusions: Solving Problems of Extrapolation

Extrapolation is a central concern for many different sciences, fields and disciplines. This thesis has investigated extrapolation in social science. In particular, the focus

has been on mechanistic extrapolation and the applicability of comparative process tracing in economics. The second chapter discussed external validity, a concept typically used in discussions of extrapolation in economics. It highlighted the different ways in which the concept is used in economics, and discussed methodological analysis by philosophers of economics. The central conclusion was that instead of understanding external validity as something inherent to causal claims, the concept should be understood as describing a relationship of generalizability or transportability between a model system and a specified target system where inferences about phenomena in the model system could be extrapolated to. When external validity is understood this way, both concepts of validity are useful in analyzing possible errors in the experimental process, at least to a certain extent.

The third chapter presented extrapolation as comparative process tracing and field experiments in economics. The chapter discussed the central concepts in understanding comparative process tracing and the key characteristics, uses of and motivations for conducting field experiments. Many argue field experiments to provide causal claims that extrapolate well. The discussion in experimental economics and philosophy of economics shows that it is not necessarily so. One solution to methodological issues of experimentation are lab-like field experiments, which combine elements from laboratory and field experiments. Last, the chapter argued that while comparative process tracing is in principle a fruitful approach for understanding extrapolation in field experiments, it is also likely to encounter challenges.

The fourth chapter consisted of a case study with two examples of studies that utilize lab-like field experiments as part of their experimental design. The examples illustrated that applying comparative process tracing to actual cases of extrapolation in economics encounters conceptual and methodological challenges. The examples show that the requirements regarding information about causal structure and background evidence, so essential for comparative process tracing, may not be possible

to attain if the relevant concept of mechanism and methodological challenges are not properly understood. Epistemological puzzles such as the extrapolator's circle and the problem of heterogeneity are an intriguing and fruitful starting point for investigating the knowledge one needs to extrapolate in a warranted way. The examples show that the epistemic warrant for an extrapolation also requires understanding the methodological problems of extrapolation.

I suggest that a general, epistemological account of extrapolation across the sciences is valuable in its own right, whether one is interested in extrapolation with information about causal mechanisms such as Steel (2008), or frameworks of analogical reasoning, such as Guala (2010) and Steel (2010). However, to understand extrapolation as an epistemological question and as a part of scientific practice, these accounts need to be complemented by a more specific methodological account of extrapolation in practice in specific fields of science. Field experiments in economics can be used for a variety of goals, from accumulating causal knowledge to testing the effectiveness of a policy. Distinguishing the epistemic and practical, epistemological and methodological issues of extrapolation relevant for each goal, and tailoring our account of extrapolation according to those goals, is a fruitful way to build a more comprehensive understanding of extrapolation in social science in general.

Finally, the examples raise questions that are relevant for future research. Extrapolation is an issue highly relevant for policy, but in the discussions on extrapolation and external validity, there are at least two kinds of claims up for extrapolation: claims about causal relations, and claims about the effectiveness of policies. Both examples in the case study provide claims about causal relationships. In contrast, the claim that the policy intervention was effective regarding academic achievement, or Steel's welfare program case, are examples of claims about policy effectiveness.

This distinction between causal claims and policy claims is made by Mireles-Flores (2016). According to Mireles-Flores, claims about causal efficacy do not

necessarily correspond to claims about policy effectiveness. Causal relations are usually studied in isolation, so that the relevant dimensions of producing the causal effect are kept under control, in order to shield the causal relata from disturbing factors. The effectiveness of policies, on the other hand, is studied by identifying the potential disturbing causes with regard to a particular policy effect. Then, a way to understand the "concurrent interactions of all the relevant causes for the production of the effect" is established. The evidence that supports a causal generalization is not necessarily the same as the evidence that supports a policy recommendation which is chosen according to the policy goal in mind. (Ibid, 25-26) Whether there are differences in extrapolating causal claims and claims about policy effectivenessn could be a relevant question when applying comparative process tracing to the social sciences.

The second question concerns the differences in extrapolation between different phases of scientific inquiry. Baetu (2015) discusses the difference between mechanistic extrapolation in basic science in medicine and mechanistic extrapolation in clinical trials. Whereas extrapolation in clinical trials can often rely on mechanistic evidence because it is available, this is typically not possible in early stages of basic research, where said mechanisms and mechanistic evidence are only being discovered and studied. (Baetu 2015, 943-944) In basic science, we may not have mechanistic information to extrapolate with, so it cannot be used to eliminate sources of error in extrapolation, as comparative process tracing instructs us to do (Ibid, 954-957). The issue is relevant to the examples discussed in this chapter, because they too are at the stage of research where mechanistic evidence to explain a phenomenon is only being discovered, construed and interpreted. This is similar to what Guala (2010) points out: Steel argues that uncertainty about causal mechanisms can hinder mechanistic extrapolation, but this may be more dependent on the stage of research, rather than the nature of the phenomenon (Guala, 2010, 1079). Whether extrapolation in the

social sciences faces the issues regarding the availability of mechanistic information at different stages of research is also a question for future research.[20]

Throughout this thesis, two lines of thinking have been discussed, side-by-side: the epistemological questions related to extrapolation, as they are typically discussed in philosophy of science, and the methodological issues of extrapolation as they are present in research in economics. Both point to the same conclusion. When one looks at the extrapolation in the social sciences more closely, the problem of extrapolation quickly transforms into multiple questions of identifying, clarifying, and controlling the epistemic risk and possibilities of error in generalizing and transporting claims from one context to another. Many issues of extrapolation are related to questions of internal validity. A solution or account that is universally applicable to "the problem of extrapolation" or "the problem of external validity" is likely to miss out on some, crucial aspect of extrapolation. Baetu's arguments about extrapolation as the taking of an epistemic risk and the implausibility of finding a universal solution to the problem of extrapolation may well hold in economics.

All in all, comparative process tracing as an account of extrapolation is theoretically comprehensive and fruitful also for the social sciences, to an extent. It shows when and how mechanistic information can act as evidence for and help guarantee the success of an extrapolation, and operates with scientific practice in mind. As an account of extrapolation in economics in practice, it has its limitations. By scrutinizing actual cases of extrapolation in experimental economics, this thesis has highlighted the benefits of understanding issues of extrapolation related to experimental methodology. This methodological understanding complements and enhances understanding of the epistemological analysis of the problem of extrapolation. Overall, many fruitful avenues of future research remain.

[20]Favereau (2016) studies the presumed analogy between randomized controlled trials in medical research and randomized field experiments in development economics more closely.

References

Alexandrova, A. (2006). Connecting economic models to the real world: Game theory and the fcc spectrum auctions. *Philosophy of the Social Sciences*, *36*(2), 173–192.

Baetu, T. M. (2015). The 'big picture': the problem of extrapolation in basic research. *The British Journal for the Philosophy of Science*, *67*(4), 941–964.

Bardsley, N., Cubitt, R., Loomes, G., Moffat, P., Starmer, C., & Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton: Princeton University Press.

Bareinboim, E., & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, *1*(1), 107–134.

Cartwright, N. (1994). *Nature's capacities and their measurement*. Oxford: Oxford University Press.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge: Cambridge University Press.

Cartwright, N. (2011). A philosopher's view of the long road from rcts to effectiveness. *The Lancet*, *377*(9775), 1400–1401.

Cartwright, N. (2012). Presidential address: Will this policy work for you? predicting effectiveness better: How philosophy helps. *Philosophy of Science*, *79*(5), 973–989.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford: Oxford University Press.

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, *92*(4), 1218–1221.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, Ill. ; Boston, Mass.: Rand McNally ; Houghton Mifflin.

Currie, A. (2015). Philosophy of science and the curse of the case study. In *The palgrave handbook of philosophical methods* (pp. 553–572). Springer.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, *4*, 3895–3962.

Favereau, J. (2016). On the analogy between field experiments in economics and clinical trials in medicine. *Journal of Economic Methodology*, *23*(2), 203–222.

Fehr, E., & Fischbacher, U. (2005). The economics of strong reciprocity. In *Moral sentiments and material interests. the foundations for cooperation in economic life* (pp. 151–193). Cambridge, MA: The MIT Press.

Finan, F., & Schechter, L. (2012). Vote-buying and reciprocity. *Econometrica*, *80*(2), 863–881.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. New York, New York: WW Norton.

Gerber, A. S., Green, D. P., & Kaplan, E. H. (2014). The illusion of learning from observational research. *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*, 9–32.

Gneezy, U., & Imas, A. (2017). Lab in the field: Measuring preferences in the wild. In *Handbook of economic field experiments* (Vol. 1, pp. 439–464). Elsevier.

Guala, F. (1999). The problem of external validity (or "parallelism") in experimental economics. *Social science information*, *38*(4), 555–573.

Guala, F. (2001). Building economic machines: The fcc auctions. *Studies in History and Philosophy of Science Part A*, *32*(3), 453–477.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of science*, *70*(5), 1195–1205.

Guala, F. (2005). *The methodology of experimental economics*. Cambridge: Cambridge University Press.

Guala, F. (2010). Extrapolation, analogy, and comparative process tracing. *Philosophy of Science*, *77*(5), 1070–1082.

Guala, F. (2012). Reciprocity: Weak or strong? what punishment experiments do (and do not) demonstrate. *Behavioral and brain sciences*, *35*(1), 1–15.

Guala, F., & Mittone, L. (2005). Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology*, *12*(4), 495–515.

Guala, F., & Mittone, L. (2010). Paradigmatic experiments: the dictator game. *The Journal of Socio-Economics*, *39*(5), 578–584.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic literature*, *42*(4), 1009–1055.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, *36*.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78.

Heukelom, F. (2011). How validity travelled to economic experimenting. *Journal of Economic Methodology*, *18*(01), 13–28.

Illari, P. M., & Williamson, J. (2012). What is a mechanism? thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119–135.

Jakiela, P. (2011). Social preferences and fairness norms as informal institutions: Experimental evidence. *American Economic Review*, *101*(3), 509–13.

Jakiela, P., Miguel, E., & Te Velde, V. L. (2015). You've earned it: estimating the impact of human capital on social preferences. *Experimental Economics*, *18*(3), 385–407.

Jiménez-Buedo, M. (2011). Conceptual tools for assessing experiments: some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology*, *18*(3), 271–282.

Jimenez-Buedo, M., & Guala, F. (2016). Artificiality, reactivity, and demand effects in experimental economics. *Philosophy of the Social Sciences*, *46*(1), 3–23.

Jimenez-Buedo, M., & Miller, L. M. (2010). Why a trade-off? the relationship between the external and internal validity of experiments. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, *25*(3), 301–321.

Khosrowi, D. (2019). Extrapolation of causal effects–hopes, assumptions, and the extrapolator's circle. *Journal of Economic Methodology*, 1–14.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, *91*(3), 437–456.

Kuorikoski, J. (2009). Two concepts of mechanism: Componential causal system and abstract form of interaction. *International Studies in the Philosophy of Science*, *23*(2), 143–160.

Levitt, S. D., & List, J. A. (2007a). On the generalizability of lab behaviour to the field. *Canadian Journal of Economics/Revue canadienne d'économique*, *40*(2), 347–370.

Levitt, S. D., & List, J. A. (2007b). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, *21*(2), 153–174.

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review*, *53*(1), 1–18.

List, J. A. (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of political Economy*, *114*(1), 1–37.

List, J. A. (2007). Field experiments: a bridge between lab and naturally occurring data. *The BE Journal of Economic Analysis & Policy*, *5*(2), 1-45.

Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, *109*(453), 25–34.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, *67*(1), 1–25.

Marcellesi, A. (2015). External validity: Is there still a problem? *Philosophy of Science*, *82*(5), 1308–1317.

Marchionni, C. (2017). Mechanisms in economics. In *The routledge handbook of mechanisms and mechanical philosophy* (pp. 423–434). Abingdon: Routledge.

Marchionni, C., & Reijula, S. (2019). What is mechanistic evidence, and why do we need it for evidence-based policy? *Studies in History and Philosophy of Science Part A*, *73*, 54–63.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, *11*(1), 57–91.

Mireles-Flores, L. (2016). *Economic science for use: causality and evidence in policy making* (Doctoral dissertation, Erasmus University Rotterdam). Retrieved from `http://hdl.handle.net/1765/93326`

Nagatsu, M., & Favereau, J. (unpublished). From the lab to the field: History and methodology of field experiments in economics.

Nicholson, D. J. (2012). The concept of mechanism in biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 152–163.

Parker, W. S. (2010). Comparative process tracing and climate change fingerprints. *Philosophy of Science*, *77*(5), 1083–1095.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Data mining workshops (icdmw), 2011 ieee 11th international conference on* (pp. 540–547).

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 579–595.

Reiss, J. (2013). *Philosophy of economics: A contemporary introduction*. Abingdon: Routledge.

Reiss, J. (2018). Against external validity. *Synthese*, 1–19.

Roe, B. E., & Just, D. R. (2009). Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics*, *91*(5), 1266–1271.

Roth, A. E. (1995). Introduction to experimental economics. In *The handbook of experimental economics, 1* (pp. 3–109). Princeton, NJ: Princeton University Press.

Santos, A. C. (2011). Experimental economics. In *The elgar companion to recent economic methodology*. Cheltenham: Edward Elgar Publishing.

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, *12*(2), 225–237.

Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of economic literature*, *43*(2), 392–436.

Steel, D. (2004). Social mechanisms and causal inference. *Philosophy of the social sciences*, *34*(1), 55–78.

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.

Steel, D. (2010). A new approach to argument by analogy: extrapolation and chain graphs. *Philosophy of Science*, *77*(5), 1058–1069.

Sugden, R. (2005). Experiments as exhibits and experiments as tests. *Journal of Economic Methodology*, *12*(2), 291–302.

Viceisza, A. C. (2016). Creating a lab in the field: economics experiments for policymaking. *Journal of Economic Surveys*, *30*(5), 835–854.

Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R., & Stuart, E. A. (2018). Target validity and the hierarchy of study designs. *American journal of epidemiology*.

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical studies*, *148*(2), 201–219.

Ylikoski, P. K. (2017). Social mechanisms. In *The routledge handbook of mechanisms and mechanical philosophy.* Routledge.